NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

MULTIVARIATE VISUALIZATION IN SOCIAL SCIENCES AND SURVEY DATA

by

William Evans

September 2013

Thesis Advisor:	Ronald D. Fricker, Jr.
Second Reader:	Samuel E. Buttrey

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** Sept 2013	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE MULTIVARIATE VISUALIZATION IN SOCIAL SCIENCES AND SURVEY DATA		**5. FUNDING NUMBERS**	
6. AUTHOR(S) William Evans			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) Department of the Navy		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number NPS.2013.0073-IR-EP7-A.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

For presentation of survey results, social science data, and other geospatial statistics requires careful attention in order to facilitate "fast and accurate" interpretation. Adding dimensionality can easily saturate the observer, leading to confusion instead of adding perspective. We produce over a dozen techniques to facilitate multivariate geospatial visualization, filter them with pilot groups, and then design a computer-based human experiment to evaluate their relative performance. In the experiment, the participants locate (with a mouse click) regions with extreme primary or secondary values and then later estimate numerically the values of these variables. We analyze these data with linear and logistic regression and general additive models to characterize the variance due to a learning effect, and then use general linear mixed-effects models to block out the variability due to individual participants and the independent and randomly-generated survey data used to generate the experiment plots. The effectiveness of a particular technique depends heavily on the goal of the presentation: a technique that provides relative perspective without distracting from the primary variable may not facilitate estimation that is as accurate as other techniques. Four scenarios are provided to qualify the presenter's intent. Only one technique performed poorly in all four scenarios and only one technique was average in all four; all remaining varied from very good to very bad between scenarios.

14. SUBJECT TERMS general linear model (GLM), general additive model (GAM), general linear mixed-effects model (GLMM), multivariate, visualization, survey, social science, geospatial, experiment, multidimensional			**15. NUMBER OF PAGES** 91
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

MULTIVARIATE VISUALIZATION IN SOCIAL SCIENCES AND SURVEY DATA

William Evans
Lieutenant Commander, U.S. Navy
B.S.E., Tulane University, 1996

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2013

Author: William Evans

Approved by: Ronald D. Fricker, Jr.
Thesis Advisor

Samuel E. Buttrey
Second Reader

Robert F. Dell
Chair, Operation Research Department

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Presentation of survey results, social science data, and other geospatial statistics requires careful attention in order to facilitate "fast and accurate" interpretation. Adding dimensionality can easily saturate the observer, leading to confusion instead of adding perspective. We produce over a dozen techniques to facilitate multivariate geospatial visualization, filter them with pilot groups, and then design a computer-based human experiment to evaluate their relative performance. In the experiment, the participants locate (with a mouse click) regions with extreme primary or secondary values and then later estimate numerically the values of these variables. We analyze these data with linear and logistic regression and general additive models to characterize the variance due to a learning effect, and then use general linear mixed-effects models to block out the variability due to individual participants and the independent and randomly-generated survey data used to generate the experiment plots. The effectiveness of a particular technique depends heavily on the goal of the presentation: a technique that provides relative perspective without distracting from the primary variable may not facilitate estimation that is as accurate as other techniques. Four scenarios are provided to qualify the presenter's intent. Only one technique performed poorly in all four scenarios and only one technique was average in all four; all remaining varied from very good to very bad between scenarios.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AIC	Akaike information criterion
ANOVA	analysis of variance
BIC	Bayesian information criterion
CED3	Center for Educational Design, Development, and Distribution
CSS	cascading style sheets
EDA	exploratory data analysis
GAM	generalized additive model
GLM	generalized linear model
HTML	hypertext markup language
IQR	inter-quartile range
NIST	National Institute of Standards and Technology
NOLH	nearly-orthogonal Latin hypercube

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Survey results, social science data, and other geospatial statistics are typically presented in briefings in univariate form, denying the observer insight into relationships with other variables or the use of descriptive measures such as confidence or uncertainty. Adding layers or dimensionality can easily saturate the graph, creating confusion instead of adding the intended perspective. More so, "confusion" is relative to each individual viewer: analysts more experienced in data exploration might quickly recognize the scope of a graph, while others might overlook the intended salient points.

In presentations including geospatial statistics, a balance must be made between rapid absorption and sufficiency of the presented data. Simpler univariate charts or plots can easily and intuitively demonstrate both the location or mean of data and their associated variability, but fail to provide areal relevance or comparison. Adding multi-dimensionality to graphs risks tipping beyond a point of tolerance or patience, resulting in misinterpretation, confusion, or dismissal.

The goal of this research is to identify graphical methods of communication that easily and effectively communicate quantitative data and their measures of uncertainty to decision makers. These are two separate dimensions, where for a graphical method to "easily communicate," we mean that it must be both intuitive and not require too much time or effort on the part of the recipient to understand, and "effective communication" translates into the numerical accuracy of the observer's interpretation.

We collaborated with Center for Educational Design, Development, and Distribution (CED3) to design over a dozen different techniques for presenting this added dimension. We brought these techniques to pilot groups to reduce the number to a more management number of techniques for an experiment. The list of candidate techniques are shown in the results plots, Figure 1. Techniques (b), (c), (f), (g), (i), and (j) survived the pilot group to be used in the experiment.

These techniques were then presented to 28 participants in a computer-based experiment (samples shown in Figure 2), asking them to first LOCATE the highest or lowest value of the average or margin of error, and then to ESTIMATE the value of a country in the same style of maps. From both of these series, we measure *time to respond* as a proxy for "easily communicate," and the error—calculated as the squared difference between their selection and the actual answer—as a proxy for "effectively communicate." We condense these results into a single binary variable that indicates whether the participant's response was "fast and accurate" or not; we group "fast and

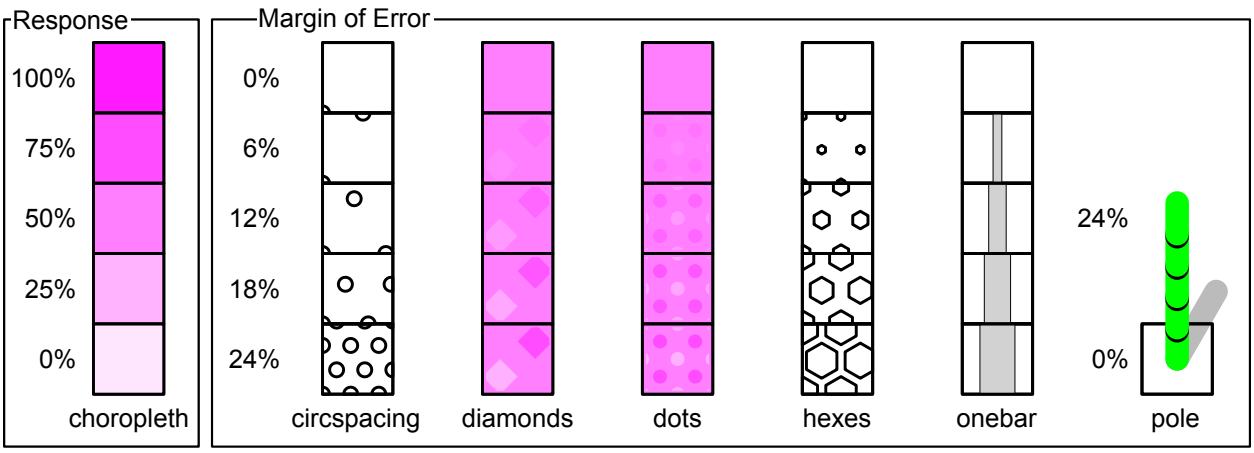


Figure 1: A colormap (choropleth) plus the six techniques that were used in the experiment. (The techniques named "diamonds" and "dots" might not display well if not in color.)

inaccurate" with "slow" regardless of accuracy since if we do not communicate the data quickly *and* accurately then the attempt was ineffective overall.

In order to properly evaluate the relative performance of each of the six techniques, we first characterize and block the sources of other variability in the data. The variances of each participant are unique, as are the variances of each randomly-generated survey data used to create the questions in the experiment. Another large source of variance in the data is in a learning curve. To properly block for this last source, we formed regression and additive models to approximate it formulaically and eventually block for it in a final mixed-effects model.

We categorize the results into two mechanisms from a mental-processing perspective: the LOCATION questions tested for recognition of relative values between countries, and the ESTIMATION

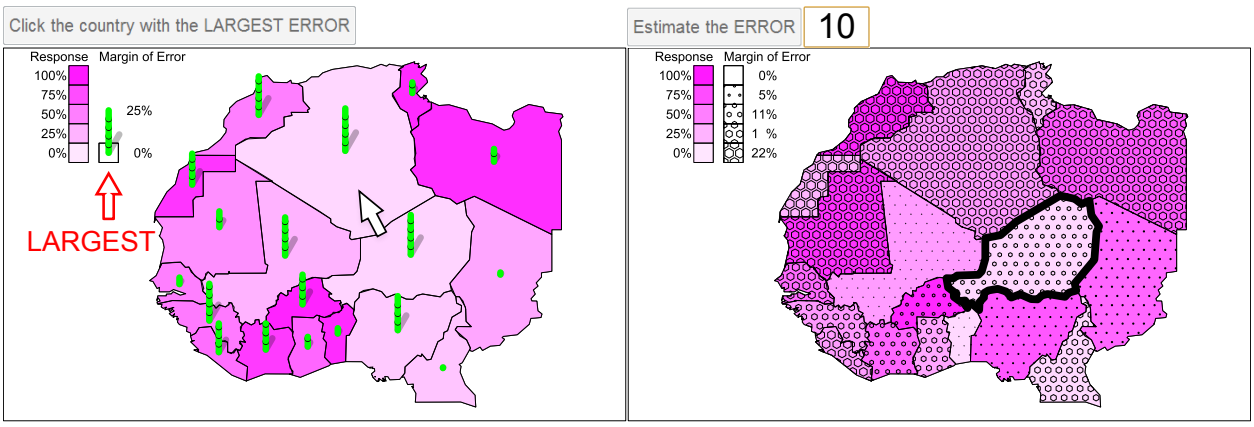


Figure 2: Sample for the LOCATION (left) and ESTIMATION (right) phases of the experiment.

questions tested for the ability to evaluate the numbers and quantify them on an absolute scale. We also split these categories into two forms of presentations: where the secondary variable (the margin of error of the country's average in this experiment) is merely informative or where it is a strong consideration. We summarize the relative performance of each of the techniques to qualify the speed of processing and error, allowing for a more informed choice of techniques. The relative performance is displayed in Figure 3.

The best performing technique is dependent on the scenario chosen. For instance, if the goal of the presentation is to display survey average responses per country and the margin of error is just informative, then the lower-left chart ("Recognize Both Variables") would indicate the *pole* technique as the best. However, if the goal is for the observers to be able to accurately quantify both variables then "Estimate Both Variables" is appropriate and shows that either of the *hexes* or *circspacing* techniques might be appropriate.

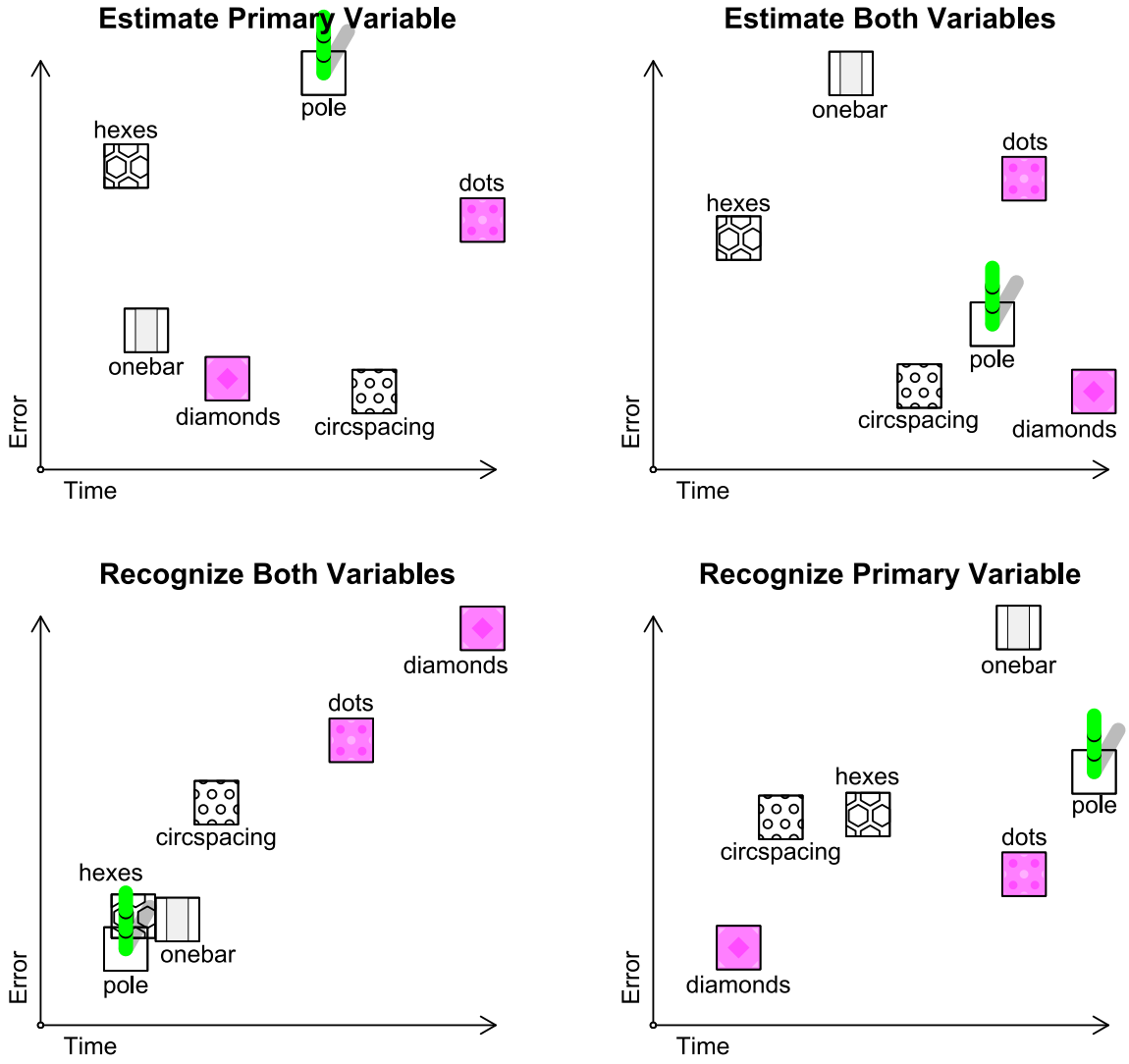


Figure 3: Results of the experiment showing relative performance of the six techniques in four presentation scenarios.

Acknowledgements

First, to my friends and colleagues in our Operations Research class: Ryan, Chuck, Matt, Karen, Roger, Ben, Kevin, Spencer, Aaron, Brad, Liz, Duncan, Chuck, Tobias, Uwe, Turgut, Begum, Scott, Jay, Rocky, Bryan, Cat, Pat, James, Matt, Aaron, Matt, Javier, Greg, Vik, Ash, Andy, and James (I think I got everybody). We are partners in our endeavors here at NPS, and I will always appreciate the time we spent together, both in and out of the classroom.

To the staff at CED3—Sherrill, Daniel, Michael—thank you for your assistance in creating the mechanism on which the rest of this experiment and this thesis relied. Without your creative perspective, I fear this would be far less interesting and insightful than it turned out.

I would like to thank Dr. Sam Buttrey for his invaluable insight and assistance. You were simultaneously direct and accommodating, never turned me away, and always provided the amount of information I needed, often answering the questions I didn't know how to ask or even that I needed to ask them.

I cannot express enough gratitude to Dr. Ron Fricker, nor can I adequately convey the effect of his motivation, teaching, and mentoring, always providing clarity and rudder at key places in my professional development and thesis writing. You enabled innovation and free thought, providing me more control over this process than I realized possible. Furthermore, you provided friendship and a sounding board in my transition from deployed military to graduate student to military retirement. I consider it a privilege to have worked with you!

Lastly, I am who I am because of the love and support of my wife, Molly. You helped me achieve this milestone in my life. You patiently allowed me to focus on the project-of-the-week/month, and always greeted me (at all hours) with a smile and a hug. You have made many sacrifices as well in this journey, and I could not have accomplished this without you. With your love and support, I feel we can accomplish anything.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Survey results, social science data, and other geospatial statistics are typically presented in briefings in univariate form, denying the observer insight into relationships with other variables or the use of descriptive measures such as confidence or uncertainty. Adding layers or dimensionality can easily saturate the graph, creating confusion instead of adding the intended perspective. More so, "confusion" is relative to each individual viewer: analysts more experienced in data exploration might quickly recognize the scope of a graph, while others might overlook the intended salient points.

Specifically, in briefings to decision makers using geospatial statistics, a balance must be made between rapid absorption and sufficiency of the presented data. Adding multi-dimensionality to graphs risks tipping beyond a point of tolerance or patience, resulting in misinterpretation, confusion, or dismissal. To understand this risk, Hyman (1953) stated that the "reaction time (to decide) seems to be a monotonically increasing function of the number of possible stimuli." The stimuli in these graphs are the variables presented, and the decision is the comprehension of the overall graph. This suggests that limiting the number of "things" in a graph – whether variables or added flair intended to highlight portions – will also minimize time necessary for interpretation for a decision. Ultimately, this form of communication can be summed up as: "a great visualization can help create a shared view of a situation and align folks on needed actions" (Sviokla 2009).

Simpler univariate charts or plots can easily and intuitively demonstrate both the location or mean of data and their associated variability. A trivial example of univariate and multivariate data in Figure 1.1 documents an experiment involving growth of guinea pig teeth. The experiment controlled the dosage of vitamin C as well as the supplement source. The left portion shows a single variable, tooth length, as a box plot and all of the contributing data points. The right portion shows the same information broken down into two contributing variables: the dose of vitamin C given along the outer x axis, and the type of supplement given on the inner x axis (per vitamin C dose). From this chart, for instance, the observer can determine that for 0.5 mg doses and supplementing with ascorbic acid (left-most yellow box plot), not only is the median in the lower portion of the interquartile range (IQR), but the whiskers are not of equal size, implying a non-uniform distribution of data points. Contrast this with the right-most box plot where the median is roughly centered in the IQR and the whiskers, though not identical, imply a less-skewed distribution. This type of display

gives an intuitive visualization for the distribution of the data and permits the observer to do fairly simple comparisons between box plots.

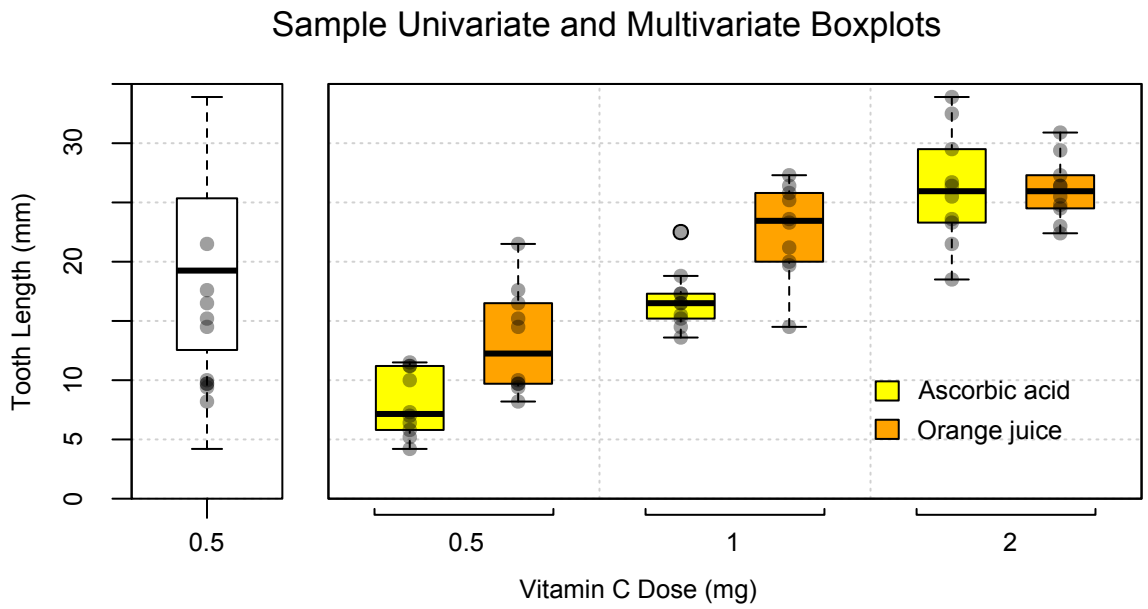


Figure 1.1: Comparison of univariate (left segment) with multivariate factored data of guinea pigs' teeth length based on supplement type and vitamin C dose. (From R Core Team (2013), `ToothGrowth` dataset.)

As simple as it appears in this example, adding a geospatial tie removes the ability to use these axes for differentiation. A common fix for complexity in geospatial results is over-simplification by omitting a measure of the variability. The assumption that the mean alone will depict the whole story is fundamentally flawed. When calculating descriptive statistics for survey results, for example, the mean for a region must be paired with the margin of error[1].

To take this example further, imagine the average response to a survey question as depicted for Mali, the center country, in Figure 1.2. The average might show roughly a 70% average response indicating a slightly-supportive populace. If the distribution of respondents' answers matched the histogram on the left of Figure 1.3 then this this might not be a problem. However, if their answers more closely matched the right chart with a bimodal distribution, despite having the same mean response as the left chart, the variability indicates a completely different response pattern to the survey question. Omission of this information (perhaps in the form of a standard deviation or margin of error measure) risks missing a key feature of the data.

[1] A margin of error is a half-width of a 95% confidence interval for the sample mean. As an example, if a survey statistic is given as "63% $\pm$ 5%", this means that the actual population average is between 58% and 68% with 95% confidence.

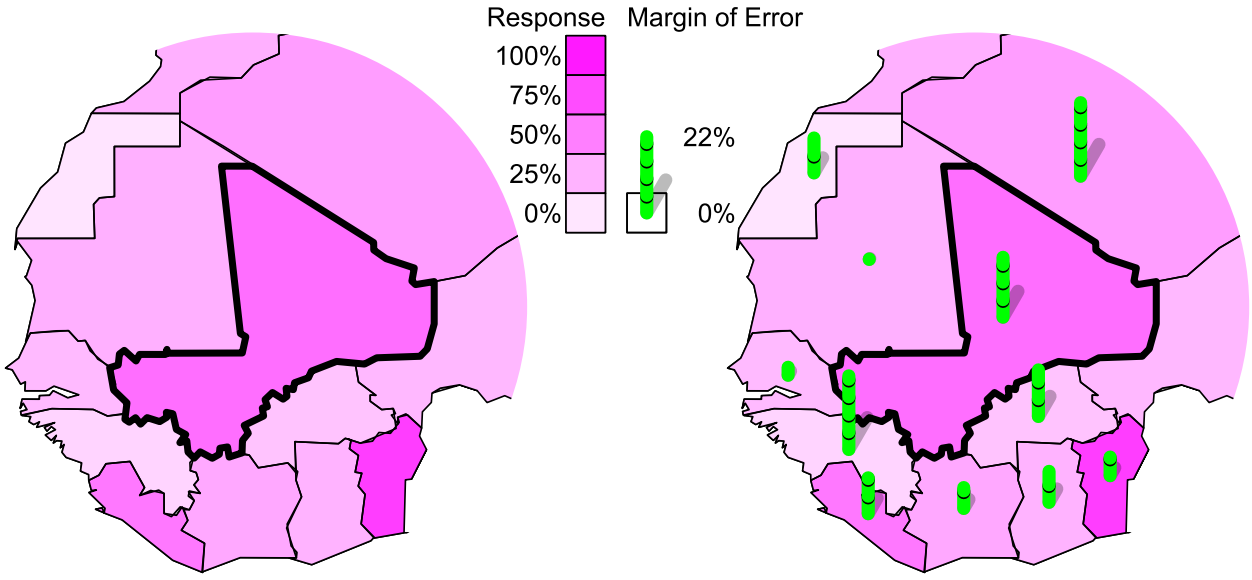


Figure 1.2: A comparison of a univariate plot (left) with one technique demonstrating a bivariate display (right). In the univariate display, it might be "obvious" to an observer that the country in the middle has the highest response and (for example) is therefore deserving of more funding and effort. However, the same data on the right also demonstrates a relatively high margin of error, which indicates that the results for Mali is highly uncertain. (In this example, each tick in the pole represents 5% of error in the estimate of the mean, so a taller pole represents more error. Though these charts are still usable in grayscale, they are designed to be read in color.)

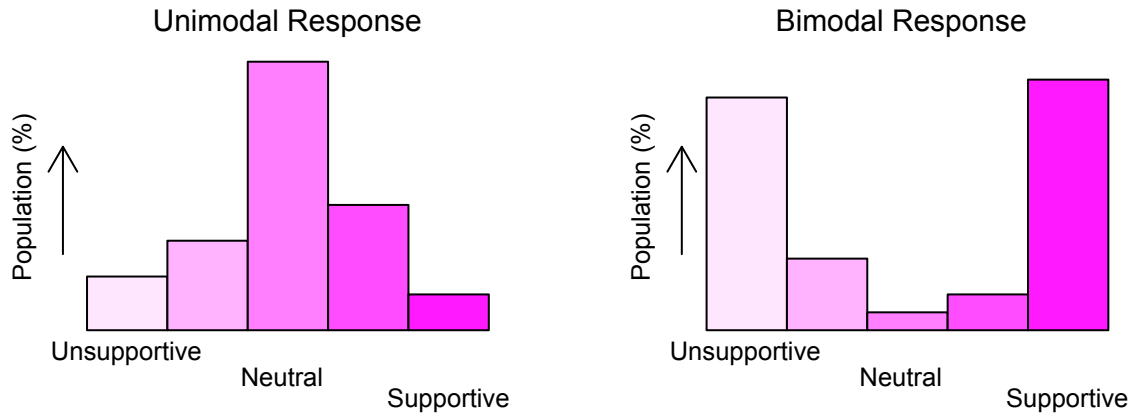


Figure 1.3: Sample histograms, both demonstrating the same mean response for the region. On the left, the majority of the country's populace is neutral. On the right, the populace is very polarized despite having the same mean response as the histogram on the left. (The colors used in these histograms match the colors used in Figure 1.2. As such, they are also designed to be read in color.)

Analysts do not have standardized techniques for displaying both the average and the variability together consistently or intuitively on geospatial plots, while non-analysts examining the plots

might not realize that it is missing. The challenge is to provide both pieces of information in such a way that the observer intuitively understands the relative importance of each variable.

1.1 Easy and Effective Graphical Communication

The goal of this research is to identify graphical methods of communication that easily and effectively communicate quantitative data (and their measures of uncertainty) to decision makers. These are two separate dimensions, where for a graphical method to "easily communicate," we mean that it must be both intuitive and not require too much effort on the part of the recipient to understand. This latter concept we refer to as the cognitive effort required by the recipient to decode and understand the graphic. For a graphic to "effectively communicate" we mean that the graphic accurately conveys the quantitative result to the recipient; that is, after looking at the graph the recipient can correctly specify or state the requisite numerical quantity.

The Merriam-Webster Dictionary defines *cognitive* as "of, relating to, being, or involving conscious intellectual activity (as thinking, reasoning, or remembering)" (Cognitive 2013). Thus, cognitive effort is the mental workload required in the conscious decoding of a graphical display. Conversely, *intuition* is defined as "the power or faculty of attaining to direct knowledge or cognition without evident rational thought and inference" (Intuition 2013). Thus, an intuitive graphic should require little cognitive effort while a non-intuitive graph should require significant cognitive effort. For the purposes of this research, we use these two words to qualitatively characterize how easy or difficult it is for a recipient to correctly interpret the information in a graphical display of data.

We note that the terms have specific technical meanings in various academic fields, ranging from psychology to artificial intelligence and education, so though volumes of literature exists, we use the terms more generally and generically. Since quantitatively measuring, for example, the actual cognitive effort required to interpret a graph in any technical sense is beyond the scope of this research, we do not find it necessary to adopt any particular precise definition.

Communication is a two-way mechanism, and the observer's skills and abilities are specifically relevant to how intuitive a particular graphical display may be. Though these tools might be useful to analysts to facilitate exploratory data analysis, for the purposes of this research, the goal is to design graphical displays that do not require specialized training or skills to correctly and easily interpret the data.

1.2 Problem Definition

As stated in the previous section, the problem this thesis addresses is the identification of graphical techniques that communicate quantitative data for easy and effective interpretation by decision makers. In particular, this research focuses on geospatial areal data that must be displayed on map and for which the display of both a statistic (for example, an average) and a measure of the uncertainty of that statistic must be displayed simultaneously.

This problem arises, for example, with survey data collected in some large region where the desire is to plot the data by sub-region. This is areal data, meaning that each statistic to be plotted corresponds with an area on the map, such as a province, state, or county. This is in contrast to point data, where each observation can be individually located on a map by its coordinates. The individual observations in areal data such as surveys can only be attributed to a region on the map, not a specific point. Furthermore, because the data to be plotted arise from a sample of the population, a measure of uncertainty is important in order to communicate to the recipient the precision of the data.

To address this problem, this research first explored a variety of techniques for displaying several types of data with experts in graphical design. These were subsequently culled to those that seemed most promising. These methods were then implemented and incorporated into a set of computer-based displays used in an experiment to measure the speed of interpretation (a proxy for *ease*) and the accuracy of value estimation (a proxy for *effective*) of multivariate areal survey results. The experiment was run with 28 participants, largely Naval Postgraduate School officer-students who are generally representative of non-analyst decision makers. The results of the experiment were then used to determine which technique provided the the best balance between ease of use and accuracy. These results were then combined into a computer-based utility allowing users to dynamically create these charts with arbitrary data, as if from a survey.

The thesis is organized as follows. Chapter 2 begins with exploration of classical data visualization techniques and adds multi-dimensionality, geospatial (areal) orientation, and combining these into one picture. Chapter 3 discusses the design of the techniques used to produce the graphical displays as well as the design of the experiments given to the participants. Chapter 4 analyzes the results of the surveys and deduces relevancy and effectiveness of each technique. Lastly, Chapter 5 lists the conclusions and potential future work for this topic.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Background and Literature Review

We begin with a review of classical data visualization, incorporating multi-dimensionality and geospatial (areal) techniques. From these sections, we provide a foundation of data-oriented graphical design for consideration in the techniques used for the experiment.

2.1 Classical Data Visualization

Data visualization, as a portion of *exploratory data analysis* (EDA) (Tukey 1977), focuses on graphical display of data for simplicity without formality or even complexity. The National Institute of Standards and Technology (NIST) describes EDA as employing mostly graphical techniques to:

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings

(National Institute of Standards and Technology 2012). Admittedly, the purpose of presentations to decision makers and viewers is not exploratory; the intent of the briefing is to highlight interesting portions of deductions or inferences. As such, and keeping in mind that the target of these briefings is often somebody not formally trained in the calculus of statistics and data analysis, NIST's presumption on useful techniques that "are graphical in nature with a few quantitative techniques" may display too much information, taking the non-analyst into a realm requiring cognitive processing and studying to interpret the data, therefore requiring more time.

Classical approaches to visualization, without geospatial ties, include bar plots, histograms, box plots, and scatterplots. These techniques provide a rich perspective on data analysis, but most of them require the user to cognitively explore each pair of variables. Cleveland discusses the prolific histogram, expanding its use via faceting (otherwise known as panels, lattice or trellis plots) but warns that "the histogram is a poor method for comparing groups of univariate measurements"

(Cleveland 1993, p. 8).

More information on classical data visualization has been described in numerous publications and is summarized by, for example, Fricker et al. (In press).

2.2 Multi-Dimensional Displays

Adding dimensionality to charts can utilize numerous techniques, not all of them geospatial or perhaps even intuitive. In bar plots, for instance, multiple variables can be shown next to each other or stacked as shown on the right side of Figure 2.1. The left side of the figure shows Yau's demonstration of Nightingale charts (Yau 2011), showing six variables on top of the categorical location. Each nominal (not geospatially-represented) location depicts six dimensions of data, the six categories of crime. In this case, location is inferred by the name of each state referenced, though effort must be expended to mentally combine the states into a comparable map. Note that he represents each of the six factors (e.g., robbery and murder) by both location in the pie chart as well as the color. In general, Tufte (1990 2001) discourages the display of more dimensions than the number of variables available. Yau's plot uses placement on the pie chart (i.e., which wedge) to indicate the crime category and reinforces this with colors, thereby using two dimensions, wedge and color, to indicate the same data variable, crime category. In this case, the combined use attempts to facilitate visual differentiation when viewing each state individually, and the picture is not implying increased relevance with exaggerated dimensionality. Geographically, visualization of this data provides ready access to each U.S. state but does not provide simple comparison or correlation by location or crime. For both charts, Yau suggests better use of white-space and explanatory text to better prepare the readers for the upcoming data: "they most likely didn't look at the data, so they might not see the same thing that you see if there's no explanation or setup" (Yau 2011, p. 330). The data itself is presented intuitively and logically, but the geospatial relationship is not easily compared.

Wickham (2009) provides an extensive package of graphing functions for the R programming language, called `ggplot2`, that provides a large number of tools for data visualization. The library simplifies representation of multiple dimensions possible by use of color (including alpha channel or transparency), size, or different panels of a plot, also referred to as facets. Figure 2.2 shows an example facet plot using an included instructional dataset *diamonds* displaying four dimensions: size, price, cut, and density. (The term facet referring to panels within a plot is not to be confused with the facets of a diamond, though perhaps the two are related.) Relatively rich dimensionality is presented here. Unfortunately, paneling the different perspectives of the data is not overlaying

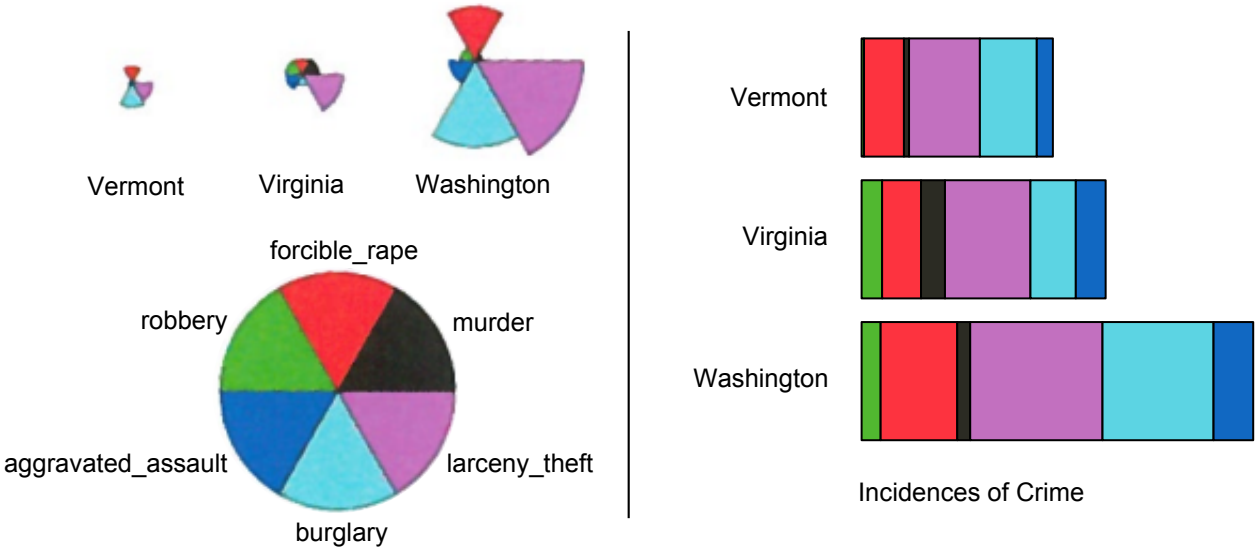


Figure 2.1: Subset of Yau's Nightingale charts showing crime in some of the United States (left). The same data is presented in stacked bar plots (right). The top-right stacked barplot does have a green bar on the left extreme, though it may not be easily discernible. (Left chart from Yau (2011, figure 7-18). The original figure included all 50 states. Viewing in color is required to differentiate several of the colors, though the horizontal stacked barplots use the same order of variables as the pie chart, starting with *robbery* and proceeding clockwise.)

two or more variables onto the same map.

Tufte states that "the time-series plot is the most frequently used form of graphical design" (Tufte 2001, p. 28), using time as the paneled variable as shown in Figure 2.3. Using limb position in the x and y axes relative to each animal, animal movement relative to the background (same axes), and time in each panel, he shows five dimensions. In order to perceive, for example, the position of the limbs on the gecko, the observer is required to examine each gecko individually, a cognitive and iterative process. Tufte asserts that these displays are usually at their best visualizing large datasets with "real variability" (Tufte 2001, p. 30).

When showing multi-dimensionality in data, Tufte (2001) warns against over-representing the magnitude of data when using area or volume to display a variable, such as in the use of bubbles or circles. By varying the radius proportionately to a variable, the perceived variable—the circle's volume—exaggerates the variability of the data. He proposes a metric called the *lie factor* which is the ratio of the size of the effect shown in the graphic and the size of the effect in the data. He suggests values of between 0.95 and 1.05 are acceptable, whereas anything outside of that show "substantial distortion, far beyond minor inaccuracies in plotting" (Tufte 2001, p. 57).

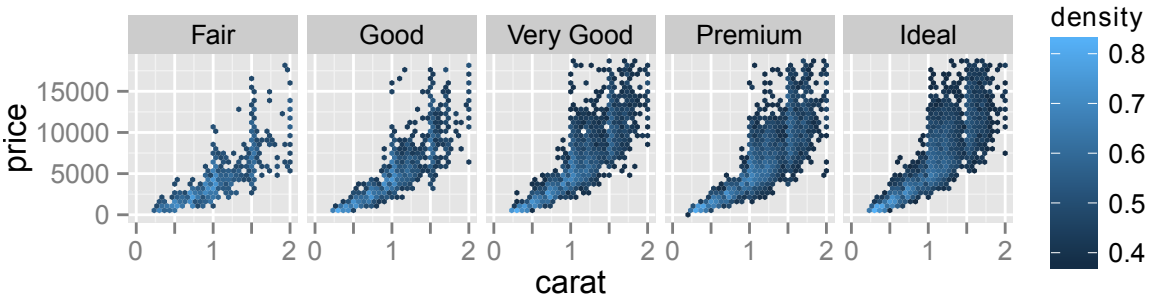


Figure 2.2: Wickham's `ggplot2` library demonstrating four dimensions of data in a panel plot. The four dimensions shown are the size (carat) and price on the x and y axes, the quality of the diamonds on the different panels, and the number of diamonds per data point (density) is depicted using color. (From Wickham (2013), density exaggerated to provide more contrast of color. Additionally, the differentiation by color is more apparent when viewed in color.)

Figure 2.3: Time-series plot of the "advance of the gecko" and the "undulations of the dorsal fin of a descending sea horse" (Tufte 2001, p. 36), showing qualitative rather than quantitative information. These use paneling as a form of multi-dimensionality.

In one format, Rosling (2013) displays multiple variables via bubbles on a scatterplot. In addition to the variables represented on the two axes, the size of the bubble represents a third variable and (optionally) bubble color represents a fourth. His software, *Gapminder World*, provides over 500 datasets for comparison; an example output is provided in Figure 2.4. This chart does show the data with a spatial reference by coding the colors based on the color of the countries, shown in the inset world map. The observer can readily see relationships between countries; for example, the dark blue bubbles representing much of Africa are predominately in the lower-left of the cluster of bubbles and there do not appear to be any large blue bubbles, in comparison to the very large light blue (India) and red (extending from southeast Asia to Australia). This method is similar to paneling, however, in that to properly compare countries, the observer must either know the colors beforehand or frequently refer to the inset map. Additionally, this method suffers when comparing more than a handful of distinct countries.

Perhaps one of the most elegant portrayals of multi-dimensional data is Charles Minard's flow map

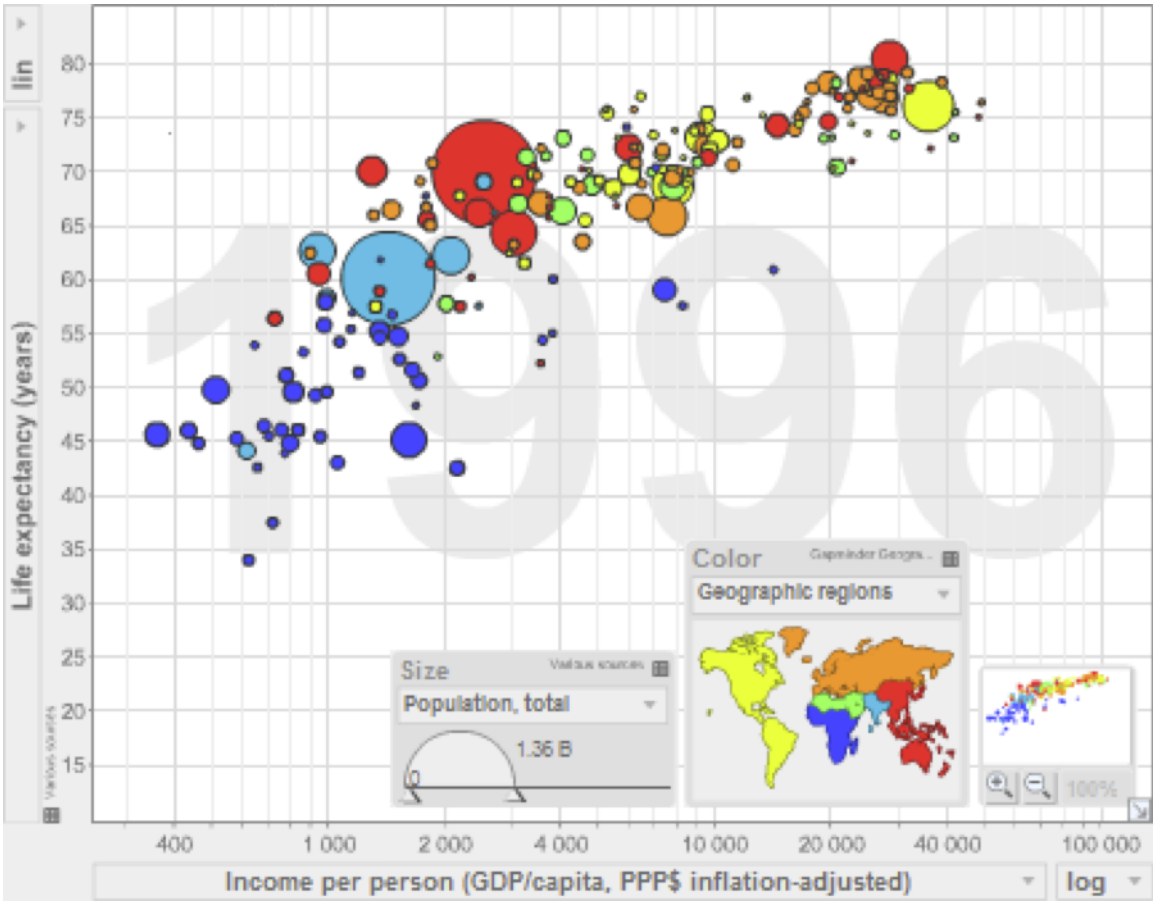


Figure 2.4: Rosling's *Gapminder World* software provides easily-interpretable displays of up to four variables, with a fifth presented in a time-series animation. Dimensions are provided by the two axes plus the bubble size and color. Data is current as of April 20, 2013. From one screen of the *Gapminder World* application (Rosling 2013). (This chart requires color to differentiate between the regions.)

of Napoleon's March to Moscow, Figure 2.5. Tufte (2001, p. 40) credits this graph as displaying six variables: "the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. ... It may well be the best statistical graphic ever drawn." This display succeeds in showing multiple dimensions while translating spatially. Instead of showing different variables for various spatial regions, though, this depicts only two variables through time: the force strength of Napoleon's army, and the temperatures to which they exposed as they move towards Moscow. The goal of this research is to expand this for multiple regions simultaneously.

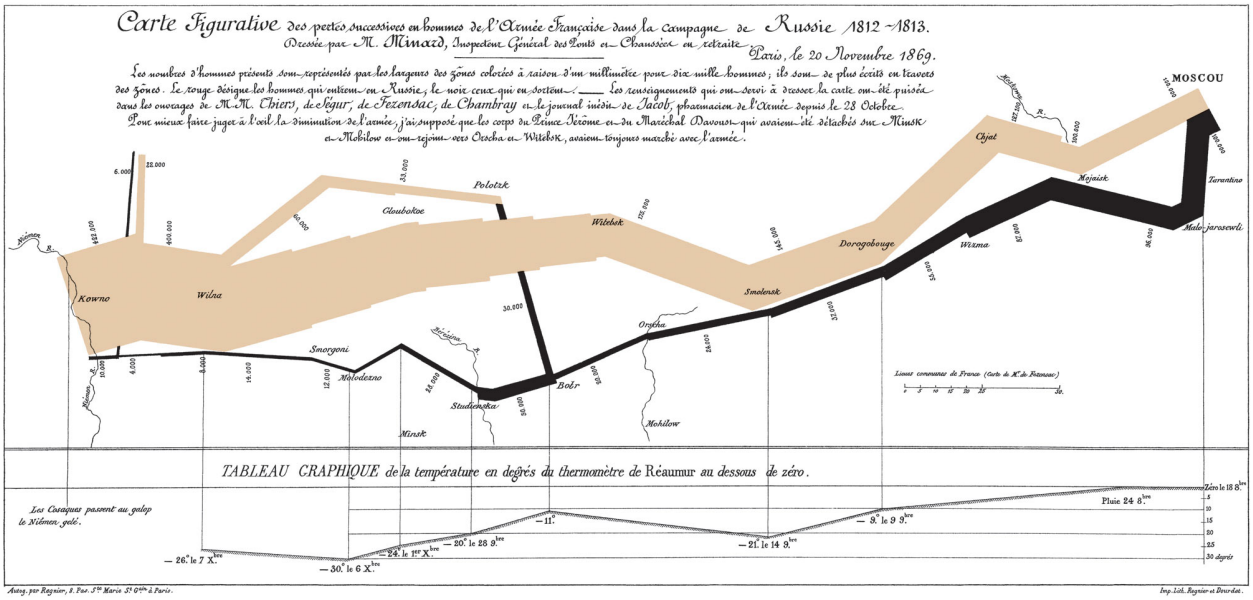


Figure 2.5: Classic graphic from Charles Minard (1781-1870) showing Napoleon's army and its progress in its assault on Moscow. This chart is both geospatial and a time-series, displaying six dimensions in the form of position and direction of movement, time-series as dates printed at various points, troop strength shown both numerically and as a function of line width, and temperature included on the bottom portion.

2.3 Geospatial Displays

Rosling's data can also be displayed geographically, as in Figure 2.6 where the two axis variables from Figure 2.4 (life expectancy and income per person) are exchanged for country centroid latitude and longitude, while the bubble size and color retain their original meanings. When compared to Figure 2.4, this facilitates spatial comparisons by overlaying data points on the map. In so doing, however, fewer data can be shown, as we no longer show the two axis variables from the chart, *income per person* and *life expectancy*. In this case, we are now showing a single variable, *total population*, by the size of the bubble. This also goes against Tufte's recommendation against displaying a variable with more than one dimension, as it uses both geographical placement of the bubbles to indicate country as well as the color to indicate which region of the world (though that helps bridge the commonality between this map and the previous non-spatial chart). Additionally, it would be easy to infer relevance of the bubble's location within a country vice understanding that, by being placed at the country's centroid, the bubble represents a datum for the country as a whole.

Yau focuses more on aesthetic presentation and design, commenting that for areal data, "Choro-

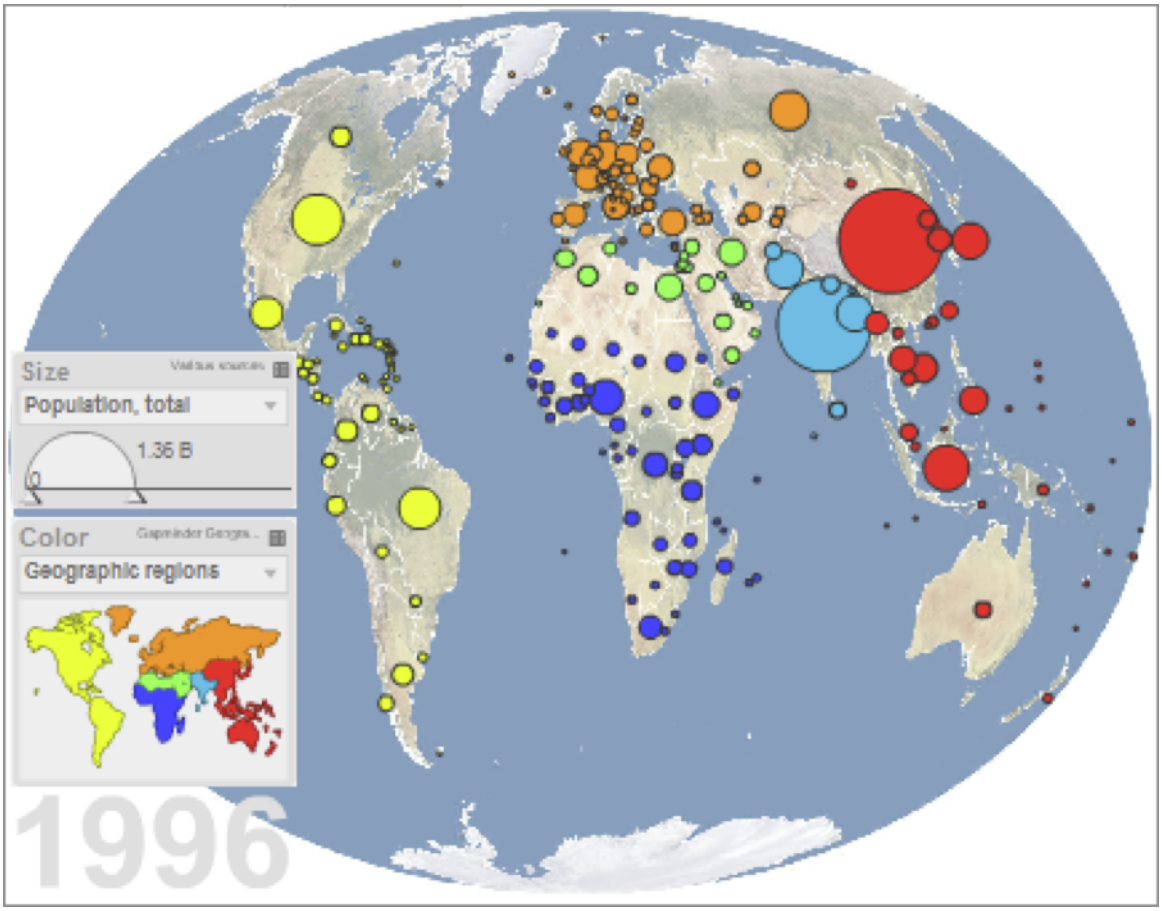


Figure 2.6: The two axis variables from Figure 2.4 are exchanged for country centroid latitude and longitude. The bubble size and color retain their relevance. Data is current as of April 20, 2013. From *Gapminder World* software display (Rosling 2013). (As this chart uses the same color coding as Figure 2.4, this also benefits from color to differentiate; however, because the colors are redundant with the geographic regions, this chart is still useful in grayscale.)

pleth maps are the most common way to map regional data" (Yau 2011, p. 286). In a manner similar to Rosling's map, Yau's display in Figure 2.7 (left chart) uses bubbles indicating Walmart store locations. The bubble size is misleading as it does not reflect the amount of stores or the size of any store: because the image in Yau's book is a snapshot of a multi-decade time-lapse animation, new stores are represented by a "blooming" bubble which quickly resorts to and remains at the fixed small size, regardless of store size or density of stores in the area. Regardless of the size, in contrast to displaying survey data, the bubbles' exact location is relevant, indicating Walmart store locations. Yau's choropleth (Figure 2.7, right chart) displays unemployment and is more relevant for regionally-defined data such as surveys.

The left method is point-specific and not applicable to surveys' regional information. The right method clearly shows regions (counties) with boundaries and shades reflecting their respective unemployment rates, though this now only shows one extra variable per region. This right chart falls victim to the original problem this paper suggests to remedy: only showing the mean response for a region and not the margin of error associated with it. If these results were based solely on a survey of a few dozen people per county, that would certainly reduce the precision of the data and the usefulness of the chart.

Wickham's `ggplot2()` can be extended with the `geom_polygon()` function that accesses map data. Combining this with the remaining paneling, binning, and other display techniques in the library allows geospatial statistics plotting, as shown in Figure 2.8. This chart is clear to understand but requires cognitive interpretation of each individual histogram in order to understand each data point by itself and in comparison with its neighbors.

Recent efforts by Livingston and Decker (2011 2012) continue with the theme of shapes and colors to expand the multidimensionality, intending to achieve up to ten variables. Figure 2.9 displays five of their techniques. These techniques are intended to enhance an analyst's ability during EDA. The techniques are intended to depict point-based data, in contrast to regional data as typically presented in survey responses. The techniques can be generalized to deal with non-point-specific regional data, though some of the techniques might imply point relevance. Livingston et al. (2011 2012 2013) continue this work by using more complex shapes, overlaid layers of techniques, and combine different techniques on the same plot for increased dimensionality. Two such examples are shown in Figure 2.10.

These techniques are certainly advanced and can provide an analyst with valuable pattern recognition and visual correlation between areas. As with an autostereogram where the image conveys depth perception to observers employing various focus techniques, Livingston's techniques require study and concentration on the chart as a whole in order to find a pattern within. For comparison, a classic autostereogram is provided in Figure 2.11. Though no studies were found that include an autostereogram for data analysis and/or presentation, it often requires significant time to study and anecdotally not everybody is able to see the embedded image.

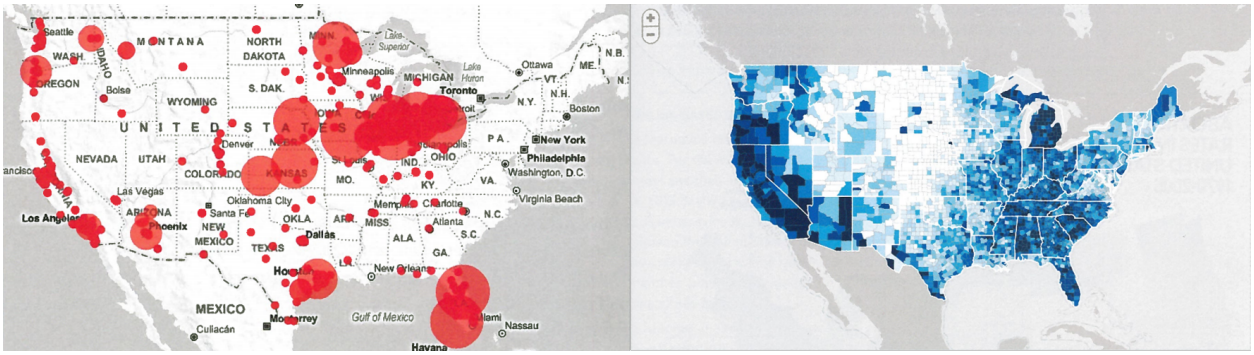


Figure 2.7: Point-based bubbles (left) display the the locations of Walmart stores at some point in the stores' history. The size of the bubble is only used to highlight, at a snapshot in time, the appearance of new stores at those locations. Choropleth map showing unemployment (right). (From Yau (2011, figures 3–25 and 3–26)).

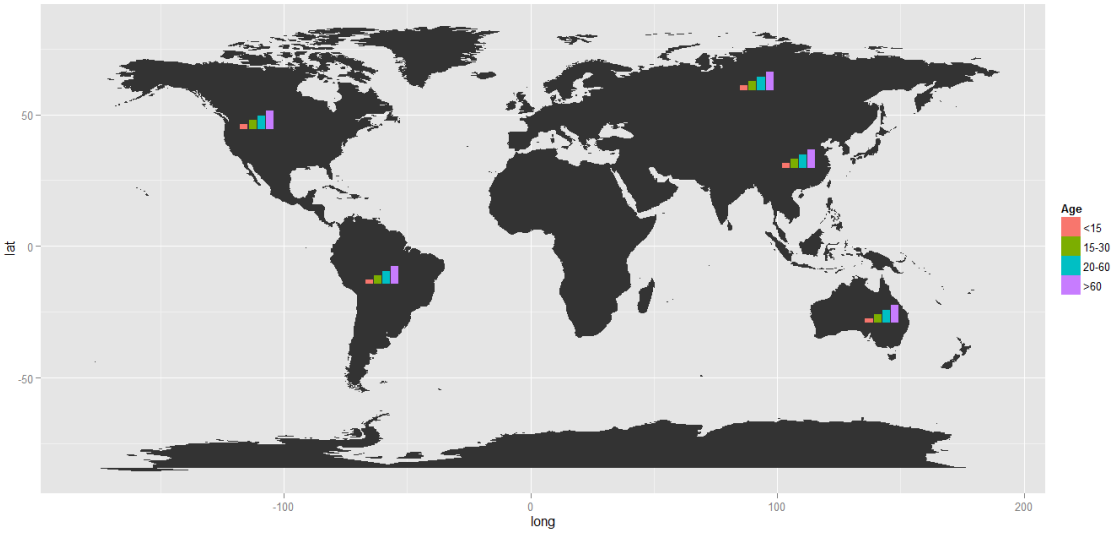


Figure 2.8: Utilizing the `ggplot2`, `ggsubplot`, and `maps` libraries of the R programming language, StackOverflow user JT85 created this as an example of mapping capabilities. Though the histograms are identical and provided solely for demonstration, if they were different then the reader would be required to "study" each histogram individually before being able to make any comparison between regions. (From StackOverflow user JT85 (2013).)

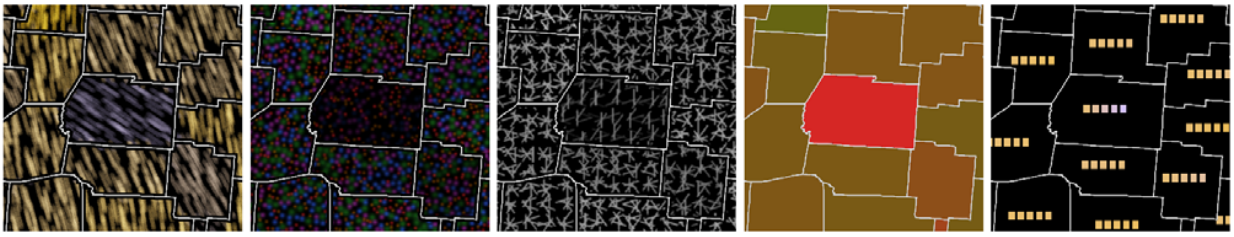


Figure 2.9: Multivariate visualization techniques evaluated in experiments by Livingston and Decker. From left to right: brush strokes, data driven spots, oriented slivers, color blending, and attribute blocks. Source: Livingston and Decker (2012). (These displays heavily rely on color.)

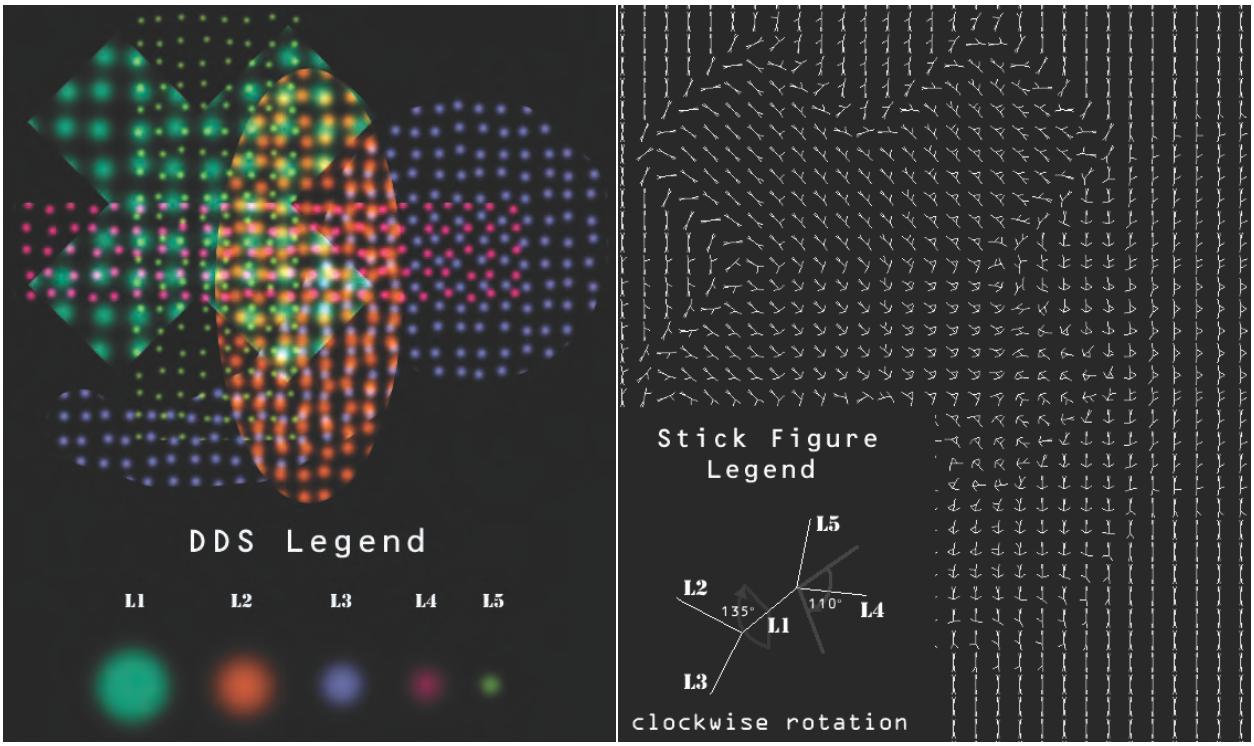


Figure 2.10: A data-driven spots technique (left) derived by combining five spot patterns, specifically configured to preclude obscuration. An abstract technique using stick figures (right) employs limbs angled with respect to the body, providing multiple data layers. This is a portion of the composite blending on the left, cropped and expanded because otherwise the stick figures would be "too small to be readable." (Extracted from Livingston et al. (2011). The left display relies heavily on color.)

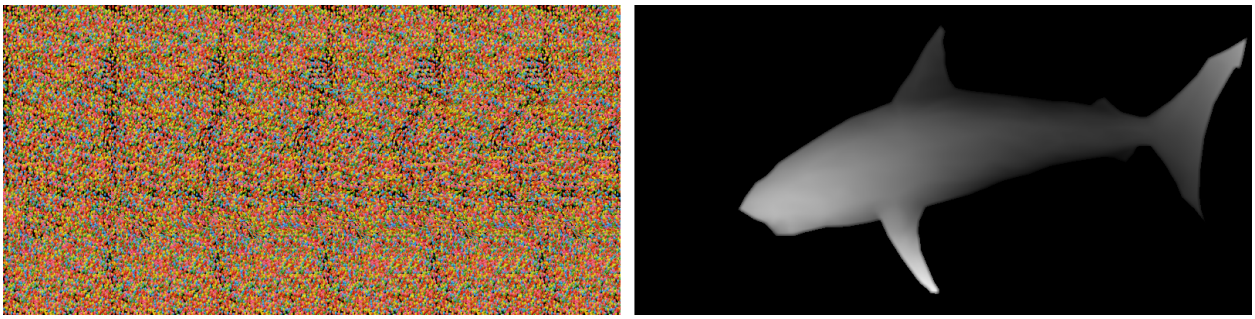


Figure 2.11: This random dot autostereogram (left) encodes a 3D scene of a shark (right) swimming before a background. This technique conveys depth perception in a 2D image. (From Hsu (2005), used under a Creative Commons Attribution-ShareAlike license, `http://creativecommons.org/licenses/by-sa/3.0/deed.en`.) Though the autostereogram on the left can still be viewed without color, depth perception might be easier to achieve with the color contrast provided by the background.)

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:
Building the Experiment

As a reminder, the purpose of this thesis is to determine techniques that provide easy and effective interpretation by non-analysts. In order to accomplish this, we first created a collection of visualization techniques for the added dimension. We employed these new techniques in survey plots and led a pilot study and an experiment to quantitatively measure the ease and effectiveness with which the participants were able to interpret the plots.

This chapter describes the decisions in building the visualization techniques and the design of the experiment. The creation of techniques followed guidance provided by graphic designers in collaboration with graphic designers in the Center for Educational Design, Development, and Distribution (CED3). These initial techniques were then implemented in the R programming language and presented to pilot groups to filter out those that performed poorly. The techniques that remained after this pilot group were included in a computer-based experiment for testing. The results from that experiment will be discussed in Chapter 4.

3.1 Creation of Techniques

The first step in testing visualization is the creation of different techniques to identify those that perform comparatively better than the others. Display variables that follow the elementary perceptual tasks of Cleveland (1994) Cleveland and McGill (1984):

- length
- direction or angle
- volume or area
- curvature
- shading, color saturation, or density

Working with the CED3, we developed display techniques that harness these characteristics. In all but one technique, we limited each technique to varying one characteristic only, both due to Tufte's warning against displaying more dimensions than available variables and so as to minimize the number of levels and factors for the design of experiments. Ten of the twelve techniques are shown in Figure 3.1.

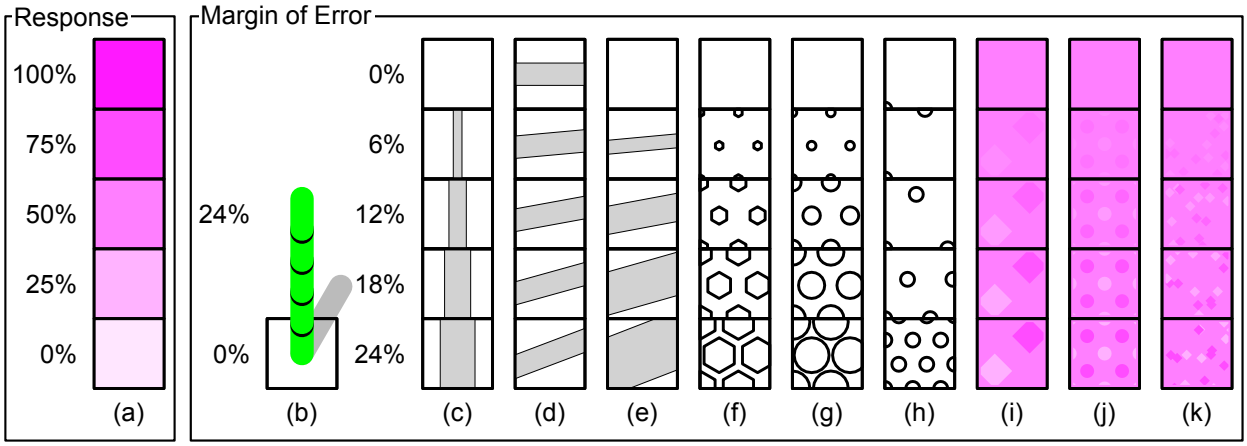


Figure 3.1: A colormap (choropleth) plus ten of the twelve techniques resulting from collaboration with the CED3. The techniques are: (a) choropleth, used solely in this experiment to represent the response to a survey question; (b) green pole; (c) vertical line; (not shown) a partial vertical line; (d) angled line; (e) angled and variable-width line; (not shown) a vertically-limited, angled, and variable-width line; (f) hexes; (g) circle spacing; (h) circle density; (i) diamonds contrast; (j) dots contrast; and (k) random dot contrast. (The last three techniques might not show well if not in color.)

The response variable was held constant through all tests, using the same technique—variable color shading—and the same colors for all tests. We kept it constant for two reasons: first, the response variable, in our case a survey question response, is likely the initial value of interest viewers look for in a chart, and since color shading is anecdotally the most common technique used for displaying survey response data, this variable will likely be the first the viewer actually "sees" (subject to interference from the secondary display technique). The second reason was to limit the number of factors and therefore the complexity of the experimental design.

The first technique for margin of error displayed, (b) the green pole, uses the length characteristic, aided by periodic black tick marks. This technique is the only one that does not fit well into a five-bin legend, as shown in the remaining techniques in Figure 3.1. There are two possible biases when using the green pole:

1. Inferred point relevance. If the poles are centered within the region or for some reason offset, the viewer could assume that the location is relevant to the data, when in fact the survey data is likely limited areally.
2. Masking of neighboring regions. Though unlikely, if one region has a particularly tall pole (a high margin of error) and a region to its north is both small and has a center point shrouded by this pole, then the northern region's pole might be partially or completely obscured.

The second technique, (c) vertical line, varies the width of a line spanning from the northern border to the southern of the region. If the margin of error is effectively zero then the bar is not visible, whereas the wider bar represents significant error. A variant of this, not shown in the figure, limits the vertical span of the line such that no region's line is any longer than any of the other regions in that chart. One possible bias for this technique is the inference of relevance of the line's length, as countered in the alternate form.

The angled line, (d), varies just the direction or angle of the line in relation to the variable. For example, a country with effectively zero margin of error would show a horizontal line and the higher margins of error might be limited to an elevation of 45° to 90°. In this example, the secondary variable, margin of error, is non-negative and will likely span from 0% to a relatively low percentage. Other examples might include variables where positive and negative values are permitted such as the relative comparison of two variables or one variable over two time period. In these examples, this angled line might have a positive or negative angle relative to the horizontal.

The angled and widened line, (e), extends (d) slightly by representing the variable both in the angle and in the line width. This violates Tufte's assertion that a chart should not have more dimensions than variables, but provides another discriminant between two similarly-angled lines. A variant of this technique, not shown in the figure, limits the vertical span of the line such that no region's line is any longer than any of the other regions in that chart, similar to the limitation on the variant for (c) vertical line.

The next two techniques, (f) hexes and (g) circles, are effectively the same, utilizing shape radius for the variable. An initial version of the hexes form varied the area. During early field testing, however, participants favored the display when the shape's size was varied by radius instead of area. This presents a bias by exaggerating the relative importance of the displayed variable, something that might be detected in the final experiment. Technique (h), circle spacing, keeps the size of the shape constant and varies the distance between them per the variable. This, too, was preferred in field testing despite the potential for bias.

The next three techniques, (i) through (k), utilize color contrast to show the possible high and low values for the given variable. The background is still based on the primary variable, so the examples depict a response of 50%. In a region with no contrast, the shapes effectively show the same color as the response, thereby rendering themselves invisible. The higher the margin of error, the more the contrast. The first and second, (i) and (j), use regularly-spaced diamonds and circles, respectively. The third, (k), randomly places dots using the two colors. There is the potential for

bias with all three of inferred point relevance; we attempt to mitigate this risk by using evenly-spaced shapes in (i) and (j) attempting to appear "too regular", and small-enough dots in (k) that they appear more as noise than as relevant points in the region.

3.2 Map Layout

For the purposes of the experiment, the layout of the plot is standardized. The response variable is always shown using the purple single-color shading. The legend is in the upper-left with five bins to indicate the variable's value (except in the case of the green pole). For all surveys, the geographic area portrayed is the northwest portion of Africa, consisting of Libya, Chad, Cameroon, and all mainland countries to the west. Only country-level survey data is generated and provided to the participants.

The goal of the experiment is to provide the user with a relatively complex map that will allow us to identify the relative ease with which a participant can recognize specific parameter extremes. As such, because of the scale of the continent, several countries were too small and would have not displayed (accurately or at all) some of the techniques. The following country pairs were combined into one shape each: Guinea and Guinea-Bissau, Seirra Leone and Liberia, and Togo and Benin.

After implementing these techniques in the R programming language, the next step was to present them to small pilot groups, with the intention of reducing the number of techniques to include in the experiment.

3.3 Pilot Study

Thirteen people were chosen to participate in the pilot study of these twelve techniques. The purpose of the pilot study was to not only reduce the number of techniques to a more manageable number (six techniques), but also to investigate possible question mechanisms for the experiment to ensure clear understanding of intent. These participants were all military officers with a variety of service and specialty backgrounds.

Each participant was given a packet that included twelve maps, showing each of the techniques. The data used to create the display was randomly generated but held constant for each participant; that is, a participant saw the same data displayed twelve different ways. Each participant within a pilot group was given a different data set.

The participants were asked to rank or score the techniques individually on a scale of at least 1 to 3, ties permitted. In other words, the participants were permitted to use any scale to compare

the techniques, as long as there were at least three levels. Responses from the participants in the pilot groups varied from 1–3 and 1–5 to a true 12-level ranking. The intent was to identify those techniques that performed particularly well and those that performed particularly poorly. The responses were then normalized on a scale from 0 to 1 and analyzed. The results are listed in Table 3.1.

Table 3.1: Normalized results of the pilot study. "Best" and "worst" columns list how many participants described that technique as being their most or least favorite, respectively. The means are generally not insightful here, possibly due to the small number of participants. The "experiment" column indicates which techniques were retained after the pilot study and used in the experiment.

Technique		**Mean**	**Std Dev**	**Best**	**Worst**	**Experiment**
(b)	Green pole	0.41	0.23	5	1	Y
(c)	Vertical line	0.81	0.27	1	7	Y
(c)	Vertical line, limited	0.77	0.27	0	6	
(d)	Angled line	0.88	0.26	1	9	
(e)	Angled / Widened line	0.81	0.31	1	7	
(e)	Angled / Widened line, limited	0.77	0.30	1	6	
(f)	Hexes	0.56	0.30	1	2	Y
(g)	Circles	0.49	0.31	2	2	
(h)	Circle spacing	0.54	0.29	2	2	Y
(i)	Diamond contrast	0.52	0.30	4	1	Y
(j)	Circle contrast	0.57	0.28	0	1	Y
(k)	Dots contrast	0.65	0.31	2	2	

The basic statistics, mean and standard deviation, provided little benefit in finding the poor performing techniques. However, despite the arbitrary ranking provided by the participants, an insightful statistic is the number of people in the pilot study who marked a particular technique their most or least favorite. This is the discriminant we will use to limit the number of techniques for the experiment. The five worst performers are both variants of (c) vertical line, (d) angled line, and both variants of (e) angled and widened line, leaving seven techniques.

In keeping with the original list of chart characteristics from Cleveland in Section 3.1, however, we opted to keep technique (c), vertical lines. Two of the contrast techniques, (f) hexes and (g) circles, are similar enough that we removed the circles. Similarly, we removed (k) random contrast dots which appears similar to (j) circle contrast. The surviving techniques are indicated in the *Experiment* column in Table 3.1.

3.4 Experiment Design

Two of the goals set forth in this thesis are to provide for *easy* and *effective* communication of data. For ease of understanding, we will use time as a proxy indicator. That is, given a map with the survey data overlaid, we measure how long it takes a participant to find the region with the desired trait. To measure the accuracy or effectiveness of a technique, we ask the participant to estimate the numerical value of a response or a margin of error.

3.4.1 Experiment Instrument

For rapid development and layout control, hypertext markup language (HTML) and cascading style sheets (CSS) are used with JavaScript code based on the jQuery library. The JavaScript code provides rudimentary information such as milliseconds between clicks and click coordinates on an image; all processing dealing with error rates and in which country the participant clicked are performed in R, outside of the experiment environment.

The first half of the test asks the participants to look for the largest or smallest value in either the response or the margin of error variable. For each question, they are shown the legend displaying the specific technique, and they are told which of the four possibilities they are being asked. This was intended to allow them to calibrate their eyes and to mitigate confusion when switching rapidly between techniques and variables. Once they have acclimated to the conditions of each question, they click on a button which begins the timer (invisibly) and populates the data on the remainder of the map. The timer for that question stops when they click on a country. To preclude ambiguity due to clicking on a border between two countries, they are advised to bias their click towards the middle of the intended country.

The second half of the test asks the participants to estimate the value of the response or the margin of error for a given country. It does this first by displaying the legend and desired variable, as well as the outline of the country to be estimated. When they are ready, the map is populated with the data and the timer again starts. It stops for that question after they have typed in a number and hit enter.

For both halves, each technique is accompanied by a brief textual description. Before the timing starts for any one question, if they are uncertain then they can either ask the proctor of the experiment or read the text. At no time during the test is there feedback on which country they clicked or the accuracy of their selection or estimation. At no time are they permitted to return to previous questions.

3.4.2 Factors and Randomization

Because the display technique for the response variable is held constant, as is the location and layout of the map components, we limited the possible factors to maximize significance in post-test analysis. We created 20 collections of random survey data; the response variable is generated from $Unif(0,1)$, and the margin of error is generated from $Unif(0,\frac{1}{4})$. The factors varied were:

- Alternate variable technique (6 levels);
- Which variable (response or margin of error);
- Which extreme (largest or smallest, first half only);
- Which country (18 levels, second half only); and
- Which survey data to use (20 levels).

The first eight participants were given 36 tests in the first half (six per technique) and 24 tests in the second half (four per technique). To test for a learning curve, these first eight were given all tests for a particular technique consecutively (in each half of the experiment). Within each technique, whether the participant was asked for the response or error, the largest or smallest, and which survey data to use, was determined randomly at runtime.

For the remaining participants, the test sequence was generated from a nearly-orthogonal Latin hypercube (NOLH) in order to minimize correlation among test columns. A NOLH does not ensure true orthogonality nor minimal correlation, but it does minimize it well. NOLHs of seven factors or less only require 17 experiments (rows within an experiment matrix). We opt for the next-larger NOLH matrix with 33 experiments in order to include all 18 countries on the map. For the first half of the experiment, we are able to reduce the maximum correlation to 6.1%; the second half is reduced to a maximum correlation of 8.1%.

This NOLH design of experiments determines which levels are to be used for each question. Since the number of questions per technique were not evenly distributed, when the experiment started the techniques were randomly ordered such that, from the NOLH table, *technique 1*, for example, did not always refer to the green pole technique. We also randomized the pairs of "response versus margin of error" variable selection and the "largest versus smallest" variable selection.

3.4.3 Participant Instruction

For the first half of the experiment, the first slide collected non-personally-identifiable demographics. The following few slides provided a basic introduction to the map, to the concept of the statistical *margin of error*, and an example of how they would proceed with the first half. Before

starting the estimation half, they were again provided two slides giving example execution.

3.5 The Experiment

Sample pages from the training and testing portions of both halves of the experiment are shown in Figures 3.2 and 3.3.

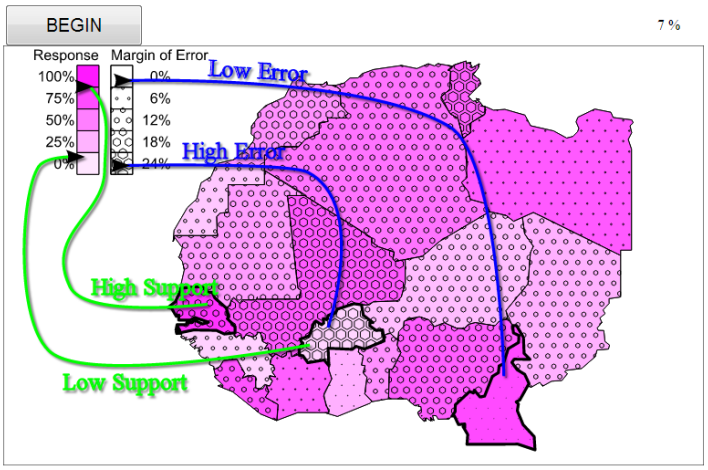
Instructions: Interpretation of All Countries

If we ask for:

- LARGEST RESPONSE (darkest color), click on Senegal (**left**-most of the three highlighted countries) with a response of 87%.
- SMALLEST RESPONSE (lightest color), click on Burkina Faso (**middle**), 20% response.
- LARGEST ERROR (largest hexes), click on Burkina Faso (**middle**), 24% margin of error.
- SMALLEST ERROR (smallest hexes), click on Cameroon (**right**-most), 0.4% margin of error.

The data and therefore the image will change for each question. You will first be given the legend for study (time for legend study is not recorded). Once you press the "READY to find ..." button, you will be shown the data and your time to find the appropriate country will start.

If you have questions, please ask them of the proctor now. When you are ready to begin the test, click BEGIN.



Variable-Height Pole

A bar where its height represents the error. The dark ticks indicate 5% intervals.

- LARGER ERROR: taller bars
- SMALLER ERROR: short bars.

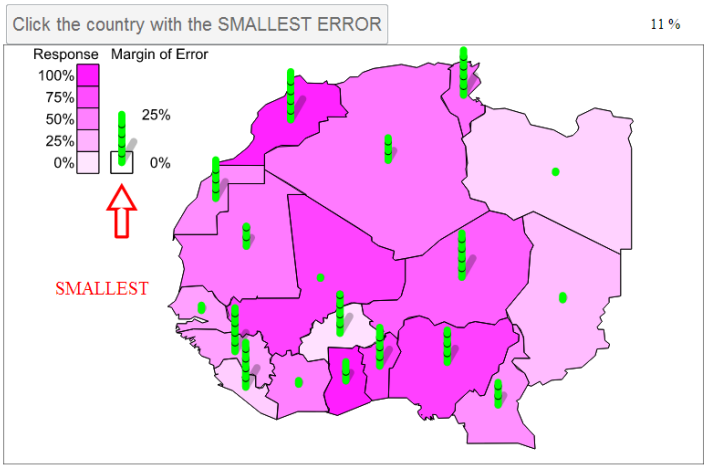


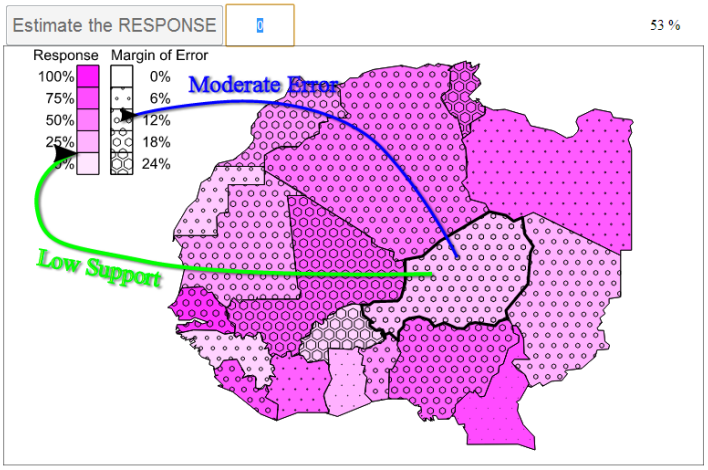
Figure 3.2: Sample pages of the training (top), and testing (bottom) portions of the first half of the experiment. (The percentage in the upper-right corner represents progress through the test.)

Instructions: Estimation of Values

You will now be asked to estimate the value of either the RESPONSE or the ERROR for a given country. The specific country will be highlighted with a thicker outline. You will be timed for this portion as well, so the sequence of clicking "READY to estimate the ..." and entering a value are the same.

Using the same example as in the first module, the country highlighted to the right (Niger) has a 19% level of support, and a 13% margin of error.

Enter "19" into the text box above the image and hit enter.



Variable-Height Pole

A bar where its height represents the error. The dark ticks indicate 5% intervals.

- LARGER ERROR: taller bars
- SMALLER ERROR: short bars.

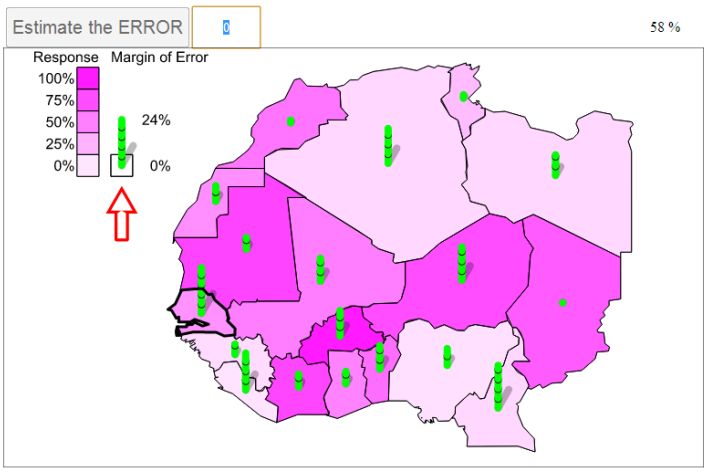


Figure 3.3: Sample pages of the training (top) and testing (bottom) portions of the second half of the experiment.

CHAPTER 4:
Experiment Analysis

This data analysis involves multiple steps, with the goal of being able to distinguish performance of each individual technique. Several of the covariates, described in Section 4.1.2, have their own variability which can easily be blocked using mixed-effects models. Since the participant can remember from question to question and therefore learn and improve his or her performance, we do not have independence among observations; they are likely tied together in a learning curve, measured in a turn-based variable that records the sequence of questions. One challenge of the data analysis is to recognize the presence of and counter the effects of this possible learning curve.

Because this thesis uses similar terminology for the experiment as for the data used to generate the experiment, we will use the following font convention:

variable: bold, indicating the variables used in the randomly-generated survey data used to generate each plot. For example, **response** and **margin of error**. All of the variables are defined in Section 4.1.2.

`variable`: fixed-width, indicating the variables used in this experiment. For example, `id`, `time`, and `correct.area`.

VARIABLE: smallcaps, indicates the dataset being referenced. Specifically, LOCATION and ESTIMATION.

The intent of the study is to find effective techniques, not necessarily perfect ones. Nor is the intent of the study to be able to predict the time required or error rate of readers of a specific technique. As such, and in keeping with the intended goal of developing a technique that facilitates easy and effective interpretation of the plot, we suggest that measuring the continuous variables of `time` and `error` are not as beneficial as simplifying the results to "fast and accurate enough" or not. More to the point, if the participant is "fast and inaccurate" or "slow" regardless of accuracy, the plot technique is ineffective in making an easy and effective communication mechanism. We quantify this binary response variable in Section 4.1.1.

To do all of this, we recognize that the challenge is to isolate the variability due to the factors in the model. For most of them, this will be handled in the final steps. However, to account for the possible learning curve, we need to characterize the time-sensitive variability before we can determine how to block it. To do that, we will first examine the measured response variables of

`time` and `error` using simple linear regression, followed by a logistic regression on the binary response variable. When we find the best-fitting transformation of the time-sequence in `seq`, we will compare its performance with a generalized additive model, specifically the smoothing function, and carry the best transformation to the last step: generalized linear mixed-effects model. In this final step, we will block for the mixed-effects and time-sequencing of `seq`, intending to reduce the background variability sufficiently to derive statistically-significant conclusions about the variance for individual techniques.

4.1 Data Description

The experiment resulted in two datasets: we call the first half the LOCATION test, where participants selected a region with the highest or lowest **response** (or **margin of error**); and we call the second half the ESTIMATION test, where they estimated the **response** or **margin** for a given country. The two datasets are not necessarily independent, but the analyses will be handled separately.

The LOCATION dataset consists of 957 observations. The first 11 participants had 36 questions, while the remaining 17 participants had the NOLH-derived 33 questions. The relevant data for each observation includes the time to respond in seconds, the country selected, and the error. (Definitions of the variables are in the next section.)

The ESTIMATION dataset consists of 825 observations. The first 11 participants had 24 questions in which they estimated one country each, and the remaining 17 participants had the NOLH-derived 33 questions. Each observation includes the time to respond, the country, and the participant's numerical estimate.

The **response** was conveyed as a choropleth, a well understood technique with which most participants were likely very familiar. Reading the **response** while potentially obscured by various secondary techniques is likely to be a different mental process than interpreting the secondary technique itself. As such, in addition to breaking the data into portions by the *test*, we also subset it by the *variable* (**response** and **margin**) he or she was instructed to interpret. For analysis of the linear regression models, we subset the data one layer further by evaluating the `time` to respond separately from the participant's `error`, as the two are not independent nor necessarily functionally correlated.

Seventy-three percent of the observations in the first half and 61% in the second half are considered good by the metric of "fast and accurate". Because we are evaluating the six techniques, a simple comparison of their performance is shown in Table 4.1. Though it is hasty to draw conclusions

from this information, two things stand out: no technique appears to dominate the others, and there may need to be a balance between time and accuracy. Figures 4.1 and 4.2 provide some comparable 95% confidence intervals for `time` and `error` (assuming independence), grouped per dataset and per technique.

Table 4.1: Percentage by variable and type of plot of correct selections or estimations for each half of the test.

Test	Variable	circspacing	diamonds	dots	hexes	onebar	pole
`Location`	Response	80.2%	84.6%	76.2%	77.8%	83.1%	77.6%
	Margin	67.6%	44.0%	48.7%	76.5%	72.4%	82.9%
`Estimation`	Response	47.1%	64.1%	50.0%	39.2%	56.9%	47.4%
	Margin	73.8%	66.7%	64.6%	79.7%	80.3%	68.4%

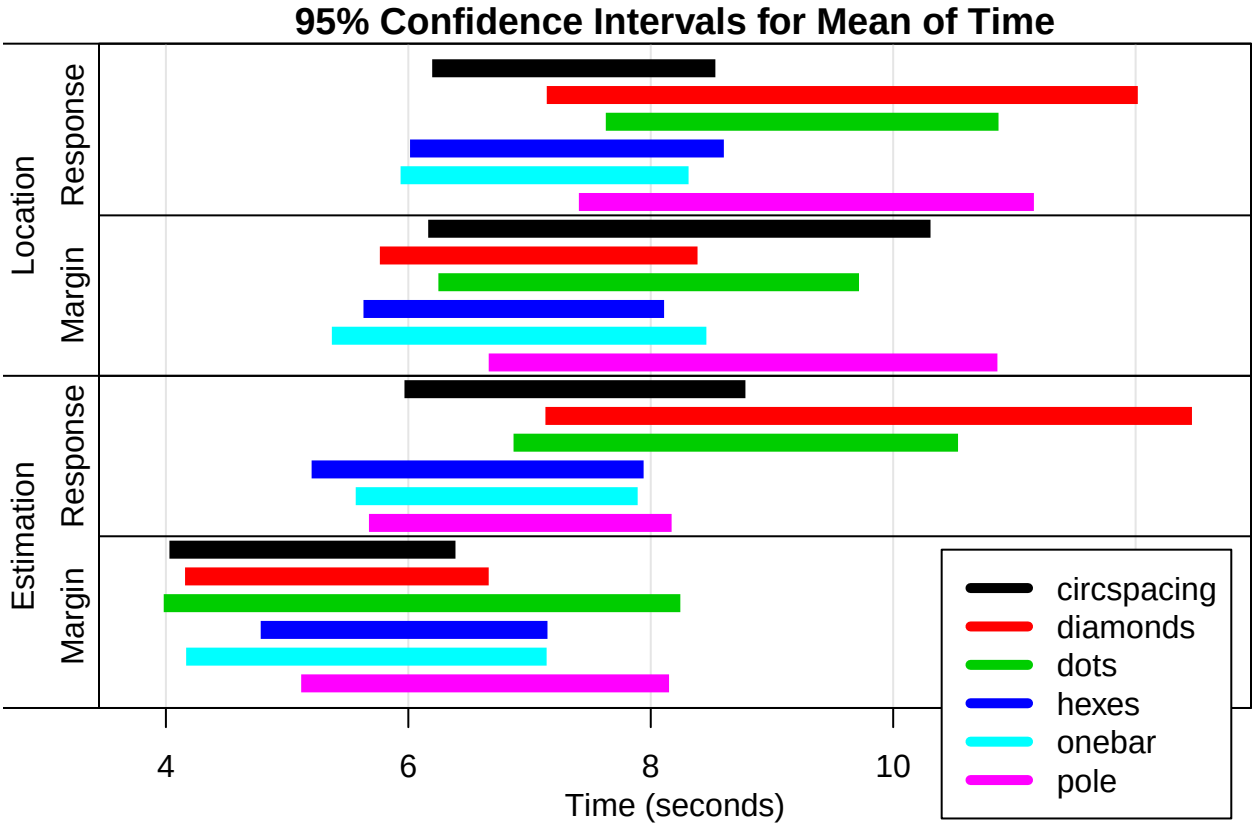


Figure 4.1: 95% confidence interval for the mean of all observations of `time`.

4.1.1 The Response Variables

`Time` is measured for each question, and begins when the data is presented and stops when the participant either clicks on a country or enters an estimate. The values range from 1.06 up to 96.10

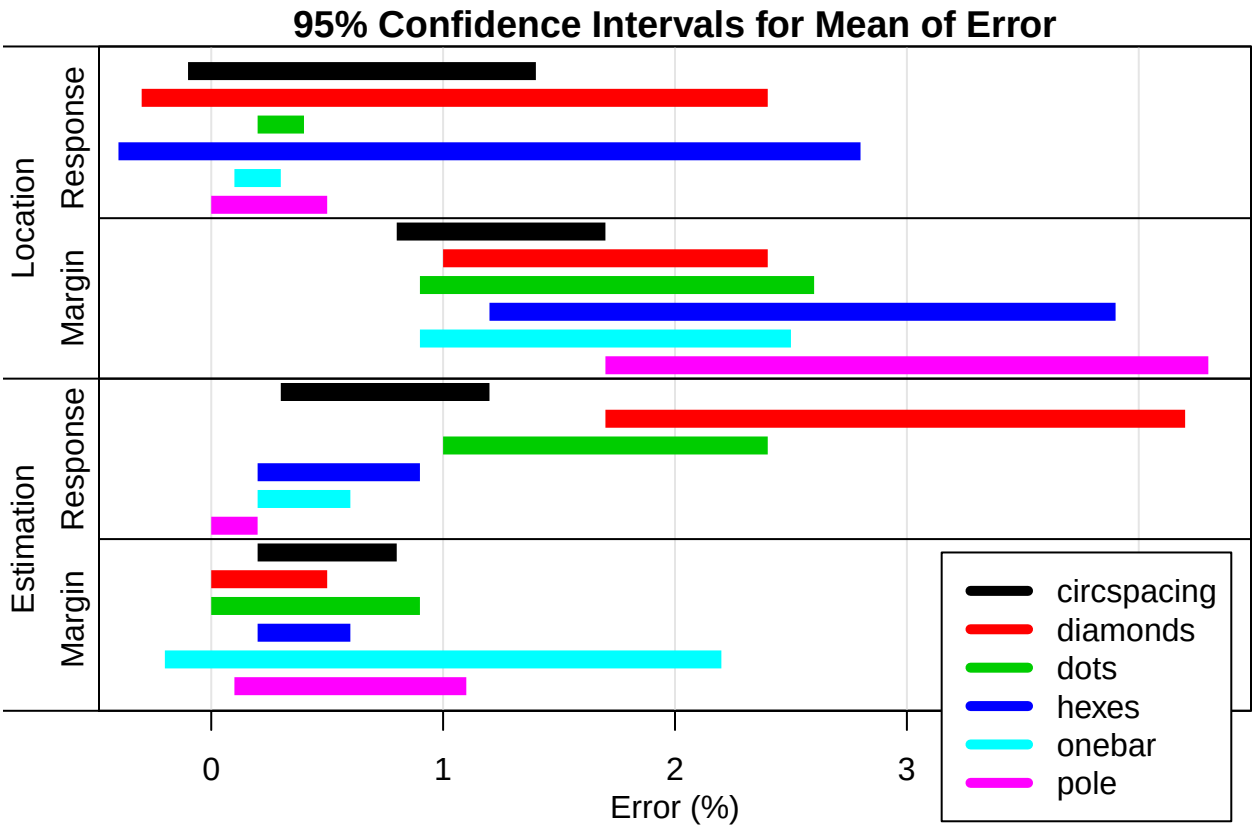


Figure 4.2: 95% confidence interval for the mean of all observations of `error`.

seconds, though the 80^{th} percentile is approximately 10 seconds. A density curve showing the lower 90% is shown in Figure 4.3, and the remaining 10% of the observations are in a decreasing right tail.

`Error` for the LOCATION test is calculated using the square of the difference between the value of the country the participant clicked on and the value of the actual "best" country. Over 50% of these observations measured 0, indicating a correct selection. The 85^{th} percentile corresponds to an error of 1%, meaning only 15% of the observations were incorrect by more than 10%. (As an example, if the **response** of the correct country is 95% and the participant clicked on a country with a value of 83%, then the difference is 12%, and the error is $0.12^2 = 0.0144$ or a 1.44% error.) For the ESTIMATION test, `error` is the square of the difference between the participant's estimate (divided by 100) and the actual value of the identified country. A 1% error corresponds to the 79^{th} percentile, meaning that the estimate provided in the remaining 21% of the observations were at least 10 percentage points incorrect. (As an example, if the correct **response** of the indicated country is 63% and the participant estimates a value of 50%, the difference is 13%, with an error

of $0.13^2 = 0.0169$, or 1.69%.) The error density curves in Figure 4.3 show the lower 90% of the observations, and the remaining 10% of the observations are in a decreasing right tail.

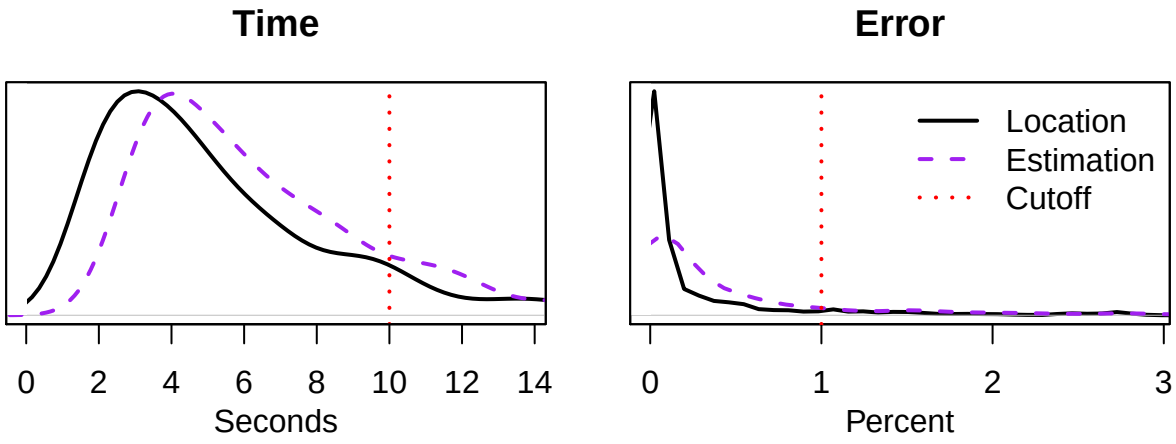


Figure 4.3: Density curves for `time` (left) and `error` (right) for the experiment. These data include at least 90% of the observations for both variables; the remaining 10%, not shown, are in decreasing right tails.

4.1.2 Covariates

The covariates in the experiment were:

`id` (categorical, 28 levels)

Each participant in the experiment was assigned a unique id.

`svy` (categorical, 20 levels)

To ensure that participants did not see the same data for all questions, 20 surveys were randomly generated. For each country used in the test, a survey response is generated from $Unif(0,1)$ and a margin of error is generated from $Unif(0,0.25)$. All observations within the fake survey are independent.

`seq` (ordinal $\in [1,36]$)

Per-participant counter for the sequences of tests. The first 11 participants received, for example, 36 questions in the LOCATION tests. Participants 12 and beyond only received 33 questions in the first half, so their `seq` will range from 1 to 33. Similarly, in the ESTIMATION test, the first 11 participants received 24 questions, so their `seq` will range from 1–24.

`tech` (categorical, 6 levels)

Indicates which technique was presented on a question. Available techniques are: circspacing, diamonds, dots, hexes, onebar, and pole, as shown in Figure 3.1.

`var` (binary)

Indicates whether the participant is told to identify the response (0) or its margin of error (1).

`minmax` (binary, LOCATION only)

Indicates whether the participant is told to find the smallest (0) or the largest (1) of whichever variable for which he or she is looking.

`click.response` (continuous $\in [0,1]$, LOCATION only)

The response of the country selected.

`click.margin` (continuous $\in [0, 0.25]$, LOCATION only)

The margin of error of the country selected.

`click.area` (continuous, LOCATION only)

The area of the country selected, in square meters.

`correct.response` (continuous $\in [0,1]$)

The correct response for the intended country, whether it be a different country (first half of the test) or the actual response for the requested country (second half). (The word "intended" means something different for each half of the test. In the LOCATION half, it means the country the participant should have clicked on, so it would have the highest or lowest value in the variable that was indicated in the question. In the ESTIMATION half, it means the country that was highlighted for the participant to estimate.)

`correct.margin` (continuous $\in [0, 0.25]$)

The correct margin of error for the intended country.

`correct.area` (continuous)

The area of the correct country. For the second half of the test, this is the same as `click.area`.

`time` (continuous)

Measured time, in seconds, between when the data is presented to the participant and a country is selected (first half) or an estimate is entered (in ESTIMATION).

`estimate` (integer $\in [0, 100]$, ESTIMATION only)

The value the participant entered as their approximation of the **response** or **margin** in a survey plot.

`error` (continuous $\in [0,1]$)

The squared difference between the value of the clicked country and the correct country (based on `var` and `minmax`) for the first half, or between the participant's estimate and the actual value (based on `var`) for the second half of the test. Because the margin of error for each survey is on a scale a quarter the size of the response variable, the error is scaled by 16 to keep the scales the same.

`RV2` (binary)

The binary response variable. A value of 1 indicates the participant's response was deemed "fast and accurate." All other responses, whether "fast and wrong" and "slow," regardless of accuracy, have a value of 0.

4.1.3 Data Cleaning

From the original dataset, there are 31 observations that we consider mistakes and adjust accordingly. Two observations from the first half of the test include mouse clicks that are not inside any of the countries. On visual observation, they are very near to a country that is easily seen as the smallest **margin** (observations #360) and smallest **response** (observation #881), so the clicks are adjusted to reflect them.

Of the remaining 29 observations, two of them it appear to have found the correct extreme (i.e., largest or smallest) but not the correct variable (i.e., **response** or **margin**). For example, if the participant is asked to locate the country with the largest **margin** and instead clicks on the country with one of the largest **responses**, the data is corrected. The criterion to discriminate a mistake from misinterpretation of the technique is a difference of greater than 0.71 for the **response** and 0.177 for the **margin**. Based on observation and post-experiment discussion with the participants, they often questioned whether they had evaluated the correct variable (or extreme) on the previous question. From this we suggest without mathematical proof it is a relatively likely mistake, therefore a cutoff error 50%. This equates to $\sqrt{0.5} \approx 0.71$ for **response** questions and $\sqrt{0.5}/4 \approx 0.177$ for **margin** questions.

As an example, the participant was instructed to locate a country with the smallest **margin** and he or she clicked on a country with a margin of error of 0.20 out of 0.25. The country with the smallest margin had a value of 0.01, so the difference is $|0.20 - 0.01| = 0.19 > 0.175$. Since that country had the smallest **response** value of 0.023, it strongly appears that the participant incorrectly selected the country with the smallest **response**, so we change `var` to reflect that the participant was looking for the response rather than the margin of error.

Similarly, 23 observations reflected a likely mistake for which extreme to select. For example, if the participant was instructed to locate the country with the *largest* margin and he or she selected the *smallest*. The same criterion is used, and the correction is to change the `minmax` variable to either "largest" or "smallest." As an example, the participant was instructed to select the country with the *largest* response, and the country selected had a response value of 0.095 and the country

with the largest response was 0.990, a difference of 0.895. The country with the smallest response had a value of 0.027 for a difference of 0.068, so `minmax` was switched from "largest" to "smallest." (Though the reader might suggest that these participants were looking for the same extreme as the previous question, only 4 of these observations followed that pattern.)

Lastly, six observations from the ESTIMATION test appeared to involve a switch in the variable.

4.2 Pre-Analysis

Based on the structure of the test and from observing the participants taking the test, we suspect that a learning curve might exist. This learning might be in one of three roles: learning the interface itself, for example getting accustomed to the layout of the page or the "prepare – click button – observe – click country" cycle; learning to look through a technique to see the underlying **response** choropleth, perhaps dealing with the technique as an obscuring or distracting element; or becoming more familiar with each technique individually.

The variable we use as an indicator of time passage through the tests is `seq`, where the values range from 1–36 (first 11 participants, LOCATION test), 1–24 (first 11 participants, ESTIMATION test), or 1–36 (all others). The `seq` values restart at 1 for the ESTIMATION test, but since we will be evaluating the LOCATION and ESTIMATION tests separately, that should not bias the results. We will be subsetting the data not just on which test was performed but additionally by which `var` the participant was instructed to find (**response** or **margin**), for a total of four models analysed.

To begin exploratory analysis, we can first examine the pairs plot to see if there are any obvious patterns or relationships. Box-Cox calculations suggest logarithmic transformations on both response variables `time` and `error` for all datasets. One of the pairs plots, shown in Figure 4.4, presents the `margin` questions from the LOCATION test. (The plots for the other three datasets can be found in Figures A.1, A.2, and A.3.) Though transforming the response variables did enable a broader view of the plots, there are no clear patterns or relationships. It might be inferred, though, that the response variables are behaving differently for each technique.

We will start model generation using linear regression and analysis of variance. The primary intent is to identify if a learning curve is suggested, and to characterize this learning as linear or non-linear. The models will suggest influence on variability by factors, but we won't know how accurate this is until we can take into account a longitudinal effect of learning as well as the mixed-effects of `id` and `svy`.

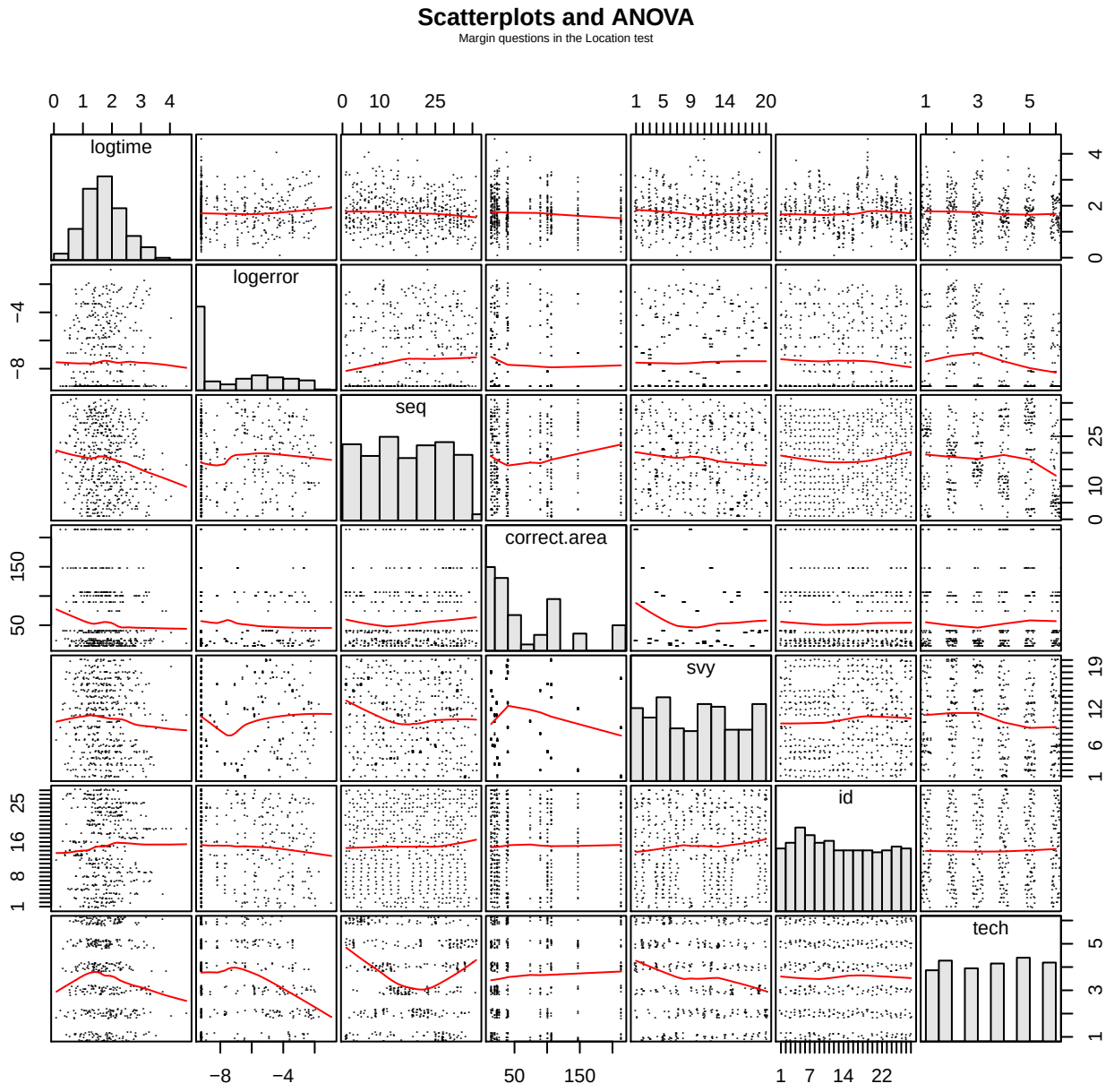


Figure 4.4: Pairs plot for `margin` questions in the LOCATION test. The `time` and `error` variables have been transformed logarithmically to help in visualization.

Next, we will use generalized additive models to see how we can approximate time progression through `seq`. Once we can compensate for it, we will use that in mixed-effect models to block for the per-participant variability in `id` and possible per-survey variability in `svy`. We contend that the remaining variability will be better conditioned to evaluate the relative performance of individual techniques.

4.3 Linear and Logistic Regression Models

4.3.1 LOCATION test, `Response` questions

We begin by reducing a main-effects linear regression to remove insignificant variables, shown in Listing 4.1. The p-value of 7.16×10^{-6} for `seq` strongly suggests that `seq` is influential on the model and that there is a learning effect. We update this model by adding $\texttt{seq}^2$ and perform an analysis of variance (ANOVA) between the two models, resulting in a p-value of 0.0078, indicating that a quadratic better approximates the passing of time in `seq`.

Listing 4.1: Linear regression of `time` in the LOCATION test, `response` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity.

```
Call:
lm(formula = I(log(time)) ~ . - error - RV2 - seq.tech - click.MARGIN -
    correct.RESPONSE - click.RESPONSE - click.area, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72382 -0.24905 -0.01727  0.26399  1.34239

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.0741523  0.1878170   5.719 2.04e-08 ***
svy2            0.1957987  0.1569165   1.248   0.2128
...svy...
svy20           0.0223180  0.1594691   0.140   0.8888
id2             0.2053930  0.1475264   1.392   0.1646
...id...
id28            0.1745416  0.1490579   1.171   0.2423
seq            -0.0118283  0.0026016  -4.547 7.16e-06 ***
techdiamonds   -0.0196130  0.0791533  -0.248   0.8044
...tech...
techpole        0.0804444  0.0812621   0.990   0.3228
minmaxSMALLEST -0.1273162  0.0545389  -2.334   0.0200 *
correct.MARGIN -0.8365239  0.5326666  -1.570   0.1171
correct.area    0.0014323  0.0006492   2.206   0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4459 on 417 degrees of freedom
Multiple R-squared:  0.6707,    Adjusted R-squared:  0.6273
F-statistic: 15.44 on 55 and 417 DF,  p-value: < 2.2e-16
```

Not surprisingly, the categorical variables show varying amounts of influence: `id` appears to be the strongest with p-values ranging from < 0.001 to 0.881; `svy` is showing mild influence ranging

from 0.057 to 0.91; and `tech` is last with mild influencers ranging from 0.21 to 0.80. The variable `minmax` appears to be a strong contributor with a p-value of 0.02, suggesting it is easier to look through the technique to a higher **response** than a lower, lighter-colored value. The presence of `correct.margin` might be interpreted as "small countries are hard to find," and `correct.area` suggests that the intended countries with smaller areas are harder to see.

The analysis of `error` is similarly performed, reduced[2], and shown in Listing 4.2. `Seq` is removed early in the reduction process, with its highest p-value before reduction of 0.62, suggesting no learning curve in making errors. We update this model by adding `seq`^{2} with a p-value of 0.731, leading us to conclude that `seq` is not a contributor to variability, so perhaps there is no learning curve in the estimation of **response**. This is not surprising since choropleths are relatively common methods of presenting univariate data and therefore no training or learning is required.

Listing 4.2: Linear regression of `error` in the LOCATION test, `response` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity.

```
Call:
lm(formula = I(log(1e-04 + error)) ~ . - time - RV2 - seq.tech -
    correct.area - seq, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9931 -1.0216 -0.3682  0.8646  9.7097

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -10.126281   2.763992  -3.664 0.000281 ***
svy2              -1.412985   0.623357  -2.267 0.023920 *
...svy...                NA         NA   0.001 0.946906
svy20              0.252546   0.594164   0.425 0.671024
id2               -0.022811   0.548593  -0.042 0.966853
...id...                 NA         NA   0.057 0.972843
id28              -0.739401   0.553947  -1.335 0.182678
techdiamonds      -0.293275   0.294452  -0.996 0.319829
...tech...               NA         NA   0.226 0.874220
techpole           0.193202   0.285414   0.677 0.498834
minmaxSMALLEST     2.156100   2.507204   0.860 0.390307
click.RESPONSE    -7.187685   1.327980  -5.412 1.05e-07 ***
click.MARGIN      -7.550888   1.320387  -5.719 2.06e-08 ***
click.area         0.003940   0.001609   2.449 0.014746 *
correct.MARGIN     6.531802   2.033600   3.212 0.001421 **
correct.RESPONSE  10.120244   3.115337   3.249 0.001254 **
---
```

[2]By *reduction*, we mean an iterative process of analyzing the model, removing the highest not-significant factor, and re-running. The categorical variables of `id`, `svy`, and `tech` were always retained.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.654 on 415 degrees of freedom
Multiple R-squared:  0.3898,    Adjusted R-squared:  0.306
F-statistic: 4.651 on 57 and 415 DF,  p-value: < 2.2e-16
```

Several of the same contributors from `time` are present in this model. New are `click.response` and `click.margin`, suggesting darker colors (for **response**) and larger or more apparent shapes/techniques (for **margin of error**) are less prone to error based on the sign of the coefficient estimates. Not surprising is `click.area`, though it is counter-intuitive that a positive coefficient suggests that larger areas lead to larger errors.

For all subsequent datasets, we reference Table 4.2, where we list the p-value associated with adding or keeping `seq`, with adding $\texttt{seq} + \texttt{seq}^2$, and with retaining the other variables. The order of the columns, though slightly different from the description in the text, allows the reader to see some patterns in significance. For instance, in the ESTIMATION test, `seq` and $\texttt{seq}^2$ are strongly significant for `time` but not at all for `error`.

Table 4.2: Summary of ANOVA p-values for the four datasets on the addition of `seq` and $\texttt{seq} + \texttt{seq}^2$ and the retention of the remaining variables. **R** is for the questions related to **response**, and **M** for **margin of error**. Gray cells are variables not present in that dataset. Empty cells are variables that start in the model and are removed for lack of significant influence.

	LOCATION Test				ESTIMATION Test			
	`time`		`error`		`time`		`error`	
Variable	**R**	**M**	**R**	**M**	**R**	**M**	**R**	**M**
`seq`	< 0.01	< 0.01	0.62	0.08	< 0.01	< 0.01	0.71	0.82
$\texttt{seq} + \texttt{seq}^2$	0.01	0.90	0.78	0.04	< 0.01	< 0.01	0.93	0.97
`estimate`							< 0.01	< 0.01
`click.area`		< 0.01	0.02					
`click.margin`		0.10	< 0.01	< 0.01				
`click.response`			< 0.01	0.03				
`correct.area`	0.03					0.17		
`correct.margin`	0.12	0.10	< 0.01	< 0.01	0.12	0.08		< 0.01
`correct.response`			0.01	0.06			< 0.01	
`minmax`	0.02	0.23	0.39	0.07				

4.3.2 LOCATION test, `Margin` questions

The analysis of the `margin` questions is very similar to the `response` questions. (The output of the linear regression can be found in Listing B.1.) Again, `seq` shows strong influence with a p-

value of 0.0008, so the presence of a learning curve is supported. The addition of $\texttt{seq}^2$ is not supported, demonstrated with a p-value of 0.896. We see the added relevance of `click.area`, `click.margin`, and `correct.margin`, suggesting that the size of the selected country as well as the degree of obscuration due to the **margin of error**. `Minmax` is retained for demonstration but we do not propose strong relevance.

In the modeling of `error`, `seq` remains in the model with a p-value of 0.076, supporting the suspicion of a learning curve in evaluating the **margin of error** of the surveys. Furthermore, addition of the quadratic improves the model with a p-value of 0.035. When compared with the analysis of `error` in Section 4.3.1, we see the addition of `click.response` and `correct.response`, suggesting the idea that the added techniques may be distracting the participant for both the **response**- and **margin**-related questions.

4.3.3 ESTIMATION test, `Response` questions

In the ESTIMATION tests, we no longer have the `minmax` factor nor any of the `click.*` factors. `Time` is highly sensitive to `seq` with a p-value of 9.4×10^{-7}, and adding $\texttt{seq}^2$ is supported with a p-value of 5.3×10^{-8}. The only other factor in the model is `correct.margin` with a p-value of 0.116. From the lack of significance of most of the variables, we can infer that the participants are relatively familiar with interpreting choropleth plots, and therefore the significance of time is related to the distraction of having the techniques overlaid.

`Error` shows no influence by `seq`, showing a p-value of 0.712 for `seq` itself and 0.933 for $\texttt{seq} + \texttt{seq}^2$. There does not appear to be a learning curve associated with estimating the response values. The estimated coefficient is negative, implying that the larger or more apparent the technique (due to a high **margin of error**), the lower the `error` in estimating it.

4.3.4 ESTIMATION test, `Margin` questions

This model is very similar to the model from `response`, showing a p-value of 1.42×10^{-7} for `seq` and 2.86×10^{-4} for $\texttt{seq} + \texttt{seq}^2$. The models also suggests influence by `correct.margin` and slightly less influence by `correct.area`, though the positive coefficient estimate would mean that larger countries correlate with longer response times.

The model for `error` is also very similar to the model of `response`, showing no significance with `seq` or $\texttt{seq} + \texttt{seq}^2$, and a strong influence from the `estimate` itself. However, with a positive coefficient estimate (unlike the negative coefficient in the model of **response**) suggests what might be less obvious, that the estimates are more accurate for lower **margins of error** in the plot.

4.3.5 Logistic Regression Modeling

Our follow-on analysis will use the binary variable `RV2` as the response, where we assign a value of 1 indicating "fast and accurate" and a 0 otherwise. Table 4.3 shows the new data.

Table 4.3: Summary of ANOVA p-values from GLMs for the four datasets on the addition of `seq` and $\texttt{seq} + \texttt{seq}^2$. Comparison is using the χ^2 test, where the null hypothesis states that there is significant reduction in deviance with the augmented model.

	LOCATION Test		ESTIMATION Test	
Variable	**Response**	**Margin**	**Response**	**Margin**
`seq`	0.023	0.764	0.707	0.003
$\texttt{seq} + \texttt{seq}^2$	0.075	0.076	0.121	0.001

A clear pattern emerges that, when using `RV2` as the response variable in the logistic regression, the quadratic of $\texttt{seq} + \texttt{seq}^2$ appears significant in all four datasets, suggesting that a quadratic might be sufficient to adequately model the learning curve through `seq`.

4.3.6 Univariate Wrap-up

Referencing Table 4.2, we again highlight some patterns that suggest further analysis. For the `time` variable, three of the four datasets show significant influence by $\texttt{seq} + \texttt{seq}^2$, and though the exception (LOCATION test, **margin** questions) seems to be a high p-value, the overall model does not suffer much with its inclusion: for the model with just `seq`, the F-statistic is 15.58 with 58 and 427 degrees of freedom, and the R-squared and adjusted R-squared are 0.6714 and 0.6284, respectively; for the model with $\texttt{seq} + \texttt{seq}^2$, the F-statistic is 15.27 with 57 and 426 degrees of freedom, and R-squared values of 0.6715 and 0.6275.

We also see significance in one of the four `error` datasets (LOCATION test, **margin** questions), and considerable lack of significance in the other three. Finally, as we shift from using continuous response variables to a binary response, we see that a quadratic of `seq` may be able to sufficiently accommodate its variability and focus more on individual techniques.

4.4 Additive Models

In the previous section, we identified where the four datasets depict presence or lack of sensitivity to `seq` in linear and non-linear forms. As an alternative to the logistic regression results, we explore additive models for possibly finding a better fit. Instead of the logistic model and its form:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \tag{4.1}$$

we use the logistic additive model:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{p} f_j\left(x_j\right) + Z\gamma, \tag{4.2}$$

where f_j are smoothing functions, Z is the matrix for variables not modeled additively, and γ are the regression parameters (Faraway 2006, p. 229-230). In this case, x_j consists of `seq`, and Z contains all other factors. As we did in the GLMs, we will use `RV2` as our binary response variable.

Looking at the results of the generalized additive model (GAM), we can visually see suggestions of patterns and assess if and how we can sufficiently approximate the smoothed `seq` by using quadratic formulæ. The remainder of the model is main-effects only, neither smoothed nor otherwise transformed and without interactions[3]. The plotted results are provided in Figure 4.5. We can see that the plot for the LOCATION test, `Response` questions (upper left plot) is largely linear, implying that loess is sufficiently handling any non-linearity that `seq` might possess. The other three plots suggest that there is still more variability in `seq`.

From an ANOVA we can reference the residual deviance and the χ^2 test to measure the loss of quality in the new transformation. We are trying to determine which type of model—loess-smoothed, linear, quadratic, or no `seq`—better handles the time-series nature of the data. In Listing 4.3, we see the ANOVA comparison between the first (loess) model and the remaining three, for the first dataset of LOCATION and `response`. These results are combined with the results from the other three datasets in Table 4.4.

Listing 4.3: Summary of a generalized additive model model comparisons using `RV2` in the LOCATION test, `response` questions. The primary model utilized `loess(seq)` and is being compared with models incorporating `seq`, `seq` + `seq`^{2}, and `seq` excluded, respectively.

```
Analysis of Deviance Table

Model 1: RV2 ~ lo(seq) + (id + svy + seq + tech + minmax + click.area +
    click.RESPONSE + click.MARGIN + correct.area + correct.RESPONSE +
    correct.MARGIN) - seq
Model 2: RV2 ~ id + svy + seq + tech + minmax + click.area + click.RESPONSE +
    click.MARGIN + correct.area + correct.RESPONSE + correct.MARGIN
  Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
1    410.69     251.08
2    413.00     252.03 -2.3064 -0.95001   0.6915
```

[3]Searching for interactions might provide predictive power or the knowledge of how specific underlying data (e.g., within the displayed survey plot) could cause problems with the display techniques. However, the data being displayed is not a controlled variable, so determining how to maximize the interpretation of the technique given specific data is not necessarily as useful.

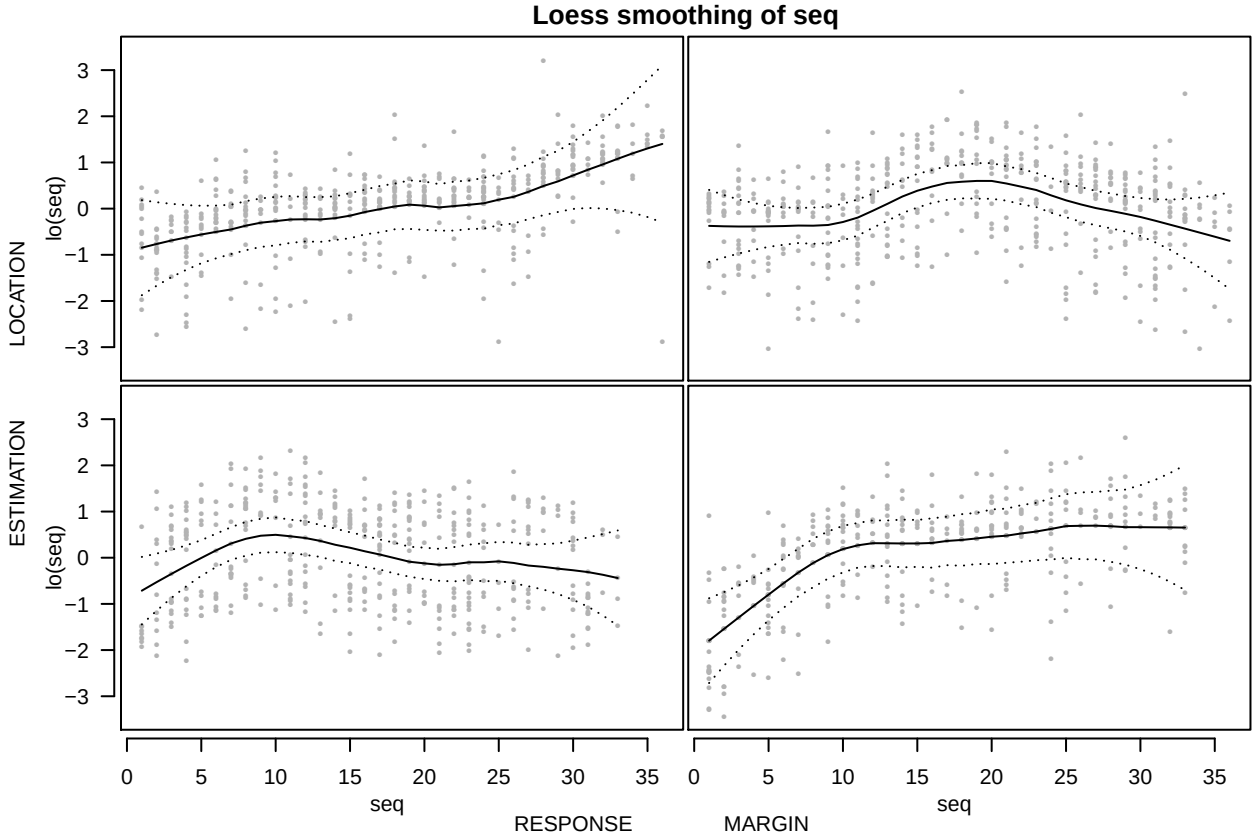


Figure 4.5: Loess smoothing of seq in each of the four datasets. A linear presentation of the loess line (upper left) might indicate an adequately-approximated seq, whereas the other non-linear presentations imply more variability in seq that smoothing does not cover.

```
Analysis of Deviance Table

Model 1: RV2 ~ lo(seq) + (id + svy + seq + tech + minmax + click.area +
    click.RESPONSE + click.MARGIN + correct.area + correct.RESPONSE +
    correct.MARGIN) - seq
Model 2: RV2 ~ id + svy + seq + tech + minmax + click.area + click.RESPONSE +
    click.MARGIN + correct.area + correct.RESPONSE + correct.MARGIN +
    I(seq^2)
  Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
1    410.69     251.08
2    412.00     252.02 -1.3064 -0.94577   0.4296

Analysis of Deviance Table

Model 1: RV2 ~ lo(seq) + (id + svy + seq + tech + minmax + click.area +
    click.RESPONSE + click.MARGIN + correct.area + correct.RESPONSE +
    correct.MARGIN) - seq
```

```
Model 2: RV2 ~ (id + svy + seq + tech + minmax + click.area + click.RESPONSE +
    click.MARGIN + correct.area + correct.RESPONSE + correct.MARGIN) -
    seq
  Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
1    410.69     251.08
2    414.00     257.20 -3.3064  -6.1204   0.1292
```

In Table 4.4, the "Resid" number on the left of each column indicates the relative change of the residuals deviance when comparing the base model of `loess(seq)` with no `seq`, `seq`, and $\texttt{seq}+\texttt{seq}^2$, respectively. This confirms that the loess is better at reducing the residual deviance, but it also shows that the quadratic ($\texttt{seq}+\texttt{seq}^2$) is closest in all three cases. The next step will be to use mixed-effects models to block for `id`, `svy`, and account for the time-series nature of `seq`, in order to focus on `tech`.

Table 4.4: Summary of ANOVA p-values from GAMs for the four datasets on removal of `loess(seq)` and the addition of `seq` and $\texttt{seq}+\texttt{seq}^2$. Comparison is using ANOVA and the χ^2 test, where the null hypothesis states that there is significant reduction in deviance with the augmented model.

	LOCATION Test				ESTIMATION Test			
	Response		**Margin**		**Response**		**Margin**	
Variable	Resid	P-value	Resid	P-value	Resid	P-value	Resid	P-value
Baseline Resid	251.08		404.82		442.91		234.65	
`seq`	252.03	0.692	415.47	0.007	452.07	0.015	241.36	0.055
$\texttt{seq}+\texttt{seq}^2$	252.02	0.430	410.39	0.027	447.99	0.039	236.49	0.283
(no `seq`)	257.20	0.129	415.56	0.017	452.22	0.034	250.20	0.002

At this point, we have two models to test for significance of between-technique differences of variation: the quadratic model derived from the GLMs, and the loess model from the GAMs. We believe the loess models will provide a better goodness of fit, but to compare these models we need to block for the mixed effects, namely `id` and `svy`.

4.5 Mixed-Effect Analysis

We believe there may be considerable variation between participants. As an example of this, see Figure 4.6. As we block for `id` in the models, one question we need to resolve is whether this variation is both in the intercept and the slope of the trends. Another way of thinking of this is that each participant might have his or her own learning curve based on prior experience as well as aptitude. To check for this, we fit models blocking for both the intercept and the slope and compare the goodness-of-fit.

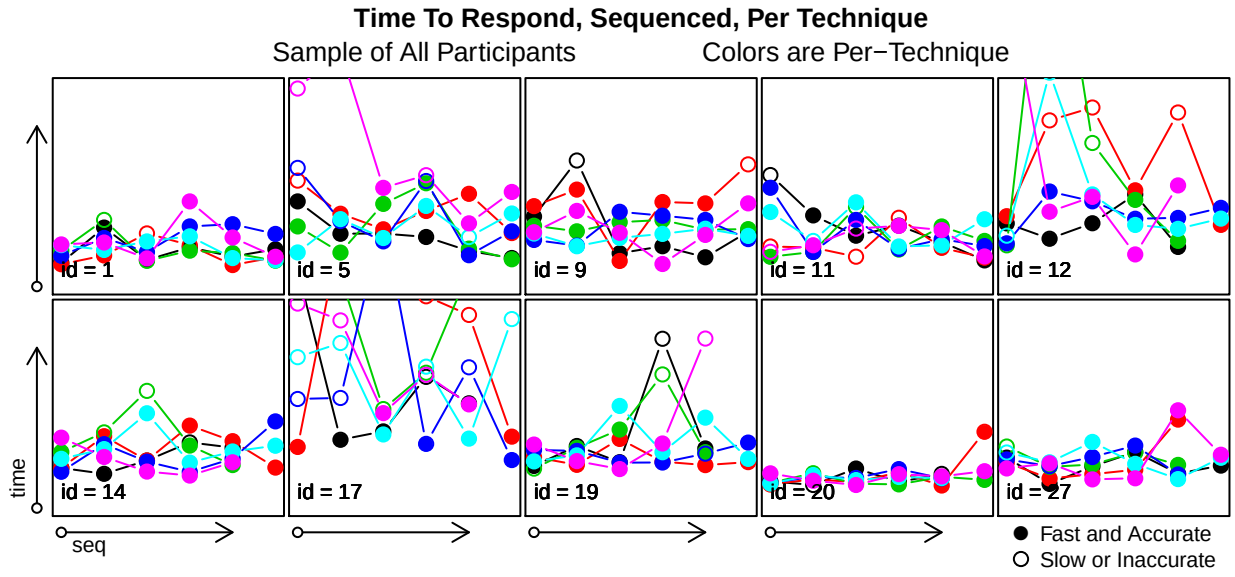


Figure 4.6: Time per sequence, colored by technique and paneled by participant. The important point to take from these plots is the difference in variability in `time` between participants. Error per sequence is provided in Figure B.1.

We have already identified the random-effects variables in the model: `id` and `svy`. For each of the four datasets, we create three models using the `lme4` package from R (Bates et al. 2013), as follows:

$$lmer\left(\texttt{RV2} \sim \texttt{loseq} + \ . \ + (1|\texttt{id}) + (1|\texttt{svy}), \ family = \text{binomial}\right) \tag{4.3}$$

$$lmer\left(\texttt{RV2} \sim \texttt{loseq} + \ . \ + (\texttt{seq}|\texttt{id}) + (1|\texttt{svy}), \ family = \text{binomial}\right) \tag{4.4}$$

$$lmer\left(\texttt{RV2} \sim \texttt{seq} + \texttt{seq}^2 + \ . \ + (1|\texttt{id}) + (1|\texttt{svy}), \ family = \text{binomial}\right) \tag{4.5}$$

$$lmer\left(\texttt{RV2} \sim \texttt{seq} + \texttt{seq}^2 + \ . \ + (\texttt{seq}|\texttt{id}) + (1|\texttt{svy}), \ family = \text{binomial}\right), \tag{4.6}$$

where `loseq` is a new variable created from the loess smoothing function on the original `seq` data, $(1|\texttt{id})$ is the notation for blocking for different intercepts among the `id` effects, and $(\texttt{seq}|\texttt{id})$ is the notation for blocking for both the intercept and slope of `id` with respect to the time-progressing variable, `seq`. Table 4.5 lists the Akaike information criterion (AIC) and Bayesian information criterion (BIC) for the four datasets.

If we adhere to the AIC for comparing model goodness-of-fit, we have strong indications to use the slightly more complicated equations (4.4) or (4.6) in the LOCATION / **response** dataset, and just equation (4.6) in ESTIMATION/**response**. The BIC, however, is dominated by the loess intercept-

Table 4.5: Comparison of AIC and BIC for the datasets and various mixed-effects models.

AIC		Location test		Estimation test	
		Response	**Margin**	**Response**	**Margin**
loseq	(1|id)	357.42	560.45	572.66	375.14
	(seq|id)	351.11	563.94	591.93	376.39
$\texttt{seq}+\texttt{seq}^2$	(1|id)	359.21	567.23	580.76	378.45
	(seq|id)	355.09	570.92	579.87	380.80

BIC		Location test		Estimation test	
		Response	**Margin**	**Response**	**Margin**
loseq	(1|id)	428.12	631.55	629.45	430.95
	(seq|id)	430.13	643.40	656.84	440.18
$\texttt{seq}+\texttt{seq}^2$	(1|id)	434.07	642.51	641.61	438.25
	(seq|id)	438.27	654.56	648.83	448.57

only model, (4.3), for all four datasets and both formulæ. If we adhere to the premise that simpler is more often correct—and because we are making several assumptions in the our analysis, this is justified—we honor the bias towards the intercept-only models.

At this point, we've controlled for as much variability as we can isolate, so we can now examine the variability within the techniques themselves. Using the intercept-only models, we perform a comparison between equations (4.3) and (4.5). A straight-forward comparison looks for lack of overlap in their confidence intervals, calculated by $\mu \pm 2\,(\sigma/\sqrt{n})$. The mixed-effects regression results are listed in Table 4.6, and the approximate 95% confidence intervals of the estimates are plotted in Figures 4.7 and 4.8.

Between the two figures, there are few differences in other than small shifts in location. This tells us that the quadratic sufficiently characterizes the relative performance of the techniques. The loess model does provide a few key differences, providing a few more statistically-significant differences.

Looking just at the loess model figure, the LOCATION/**margin** dataset shows considerable performance difference, specifically that diamonds and dots clearly perform below the rest, and hexes and pole perform significantly better than three of the four others. In contrast, for the ESTIMATION/**response** dataset places diamonds above dots and hexes. And ESTIMATION/svymargin puts hexes above diamonds, dots, and pole.

Table 4.6: Regression coefficients for `tech` variables for both the loess-based and quadratic-based models. The technique "circspacing" is the baseline level from which the others are compared, which is why its estimate is 0 for all models.

			Loess		Quadratic	
TEST	**Question**	`Tech`	Estimate	StdErr	Estimate	StdErr
LOCATION	**Response**	circspacing	0.000	0.055	0.000	0.056
		diamonds	0.057	0.055	0.060	0.056
		dots	-0.017	0.055	-0.009	0.055
		hexes	-0.021	0.056	-0.016	0.056
		onebar	-0.009	0.057	0.008	0.057
		pole	0.015	0.056	0.026	0.058
	Margin	circspacing	0.000	0.069	0.000	0.070
		diamonds	-0.216	0.068	-0.202	0.068
		dots	-0.177	0.071	-0.156	0.071
		hexes	0.140	0.067	0.122	0.067
		onebar	0.101	0.066	0.096	0.068
		pole	0.207	0.067	0.200	0.069
ESTIMATION	**Response**	circspacing	0.000	0.080	0.000	0.081
		diamonds	0.128	0.077	0.127	0.078
		dots	-0.094	0.081	-0.068	0.080
		hexes	-0.083	0.079	-0.045	0.080
		onebar	0.060	0.079	0.068	0.081
		pole	0.027	0.078	-0.007	0.080
	Margin	circspacing	0.000	0.065	0.000	0.066
		diamonds	-0.023	0.065	-0.022	0.065
		dots	-0.074	0.066	-0.060	0.066
		hexes	0.110	0.064	0.114	0.064
		onebar	0.027	0.064	0.029	0.064
		pole	-0.029	0.066	-0.025	0.067

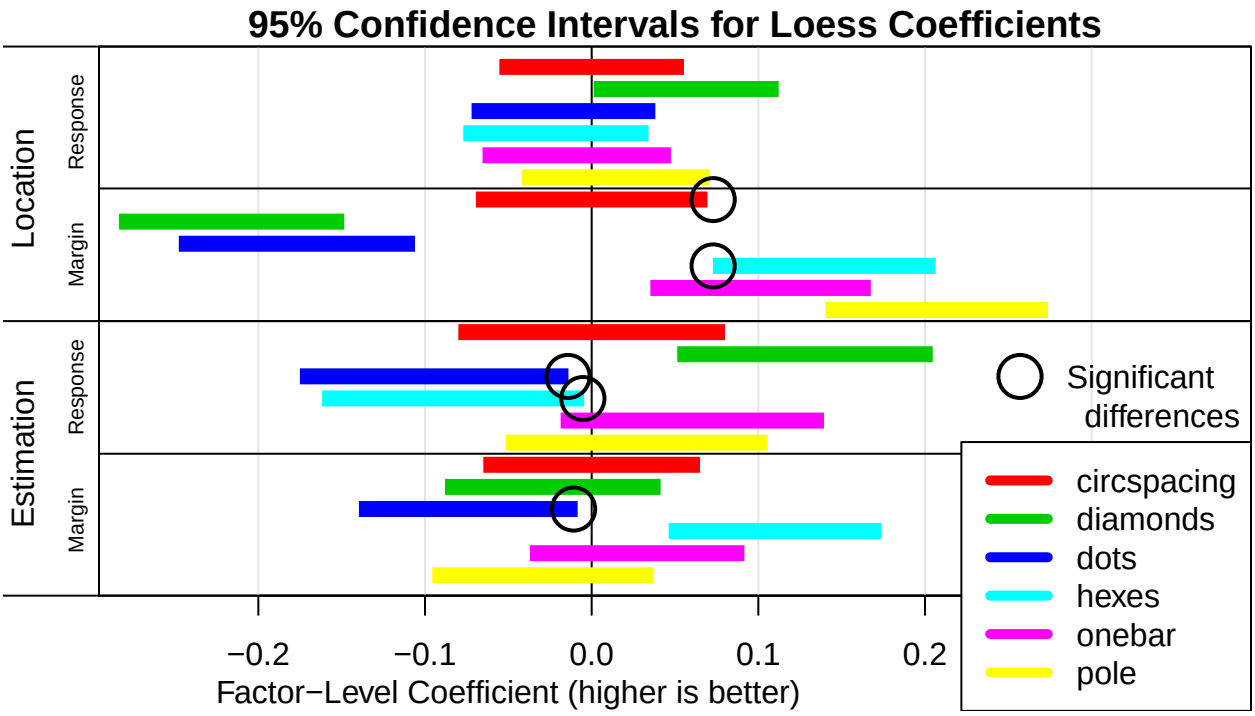


Figure 4.7: Confidence intervals the coefficient estimates in the loess-based models. The highlighted "significant differences" are between the loess and the quadratic coefficients for the techniques.

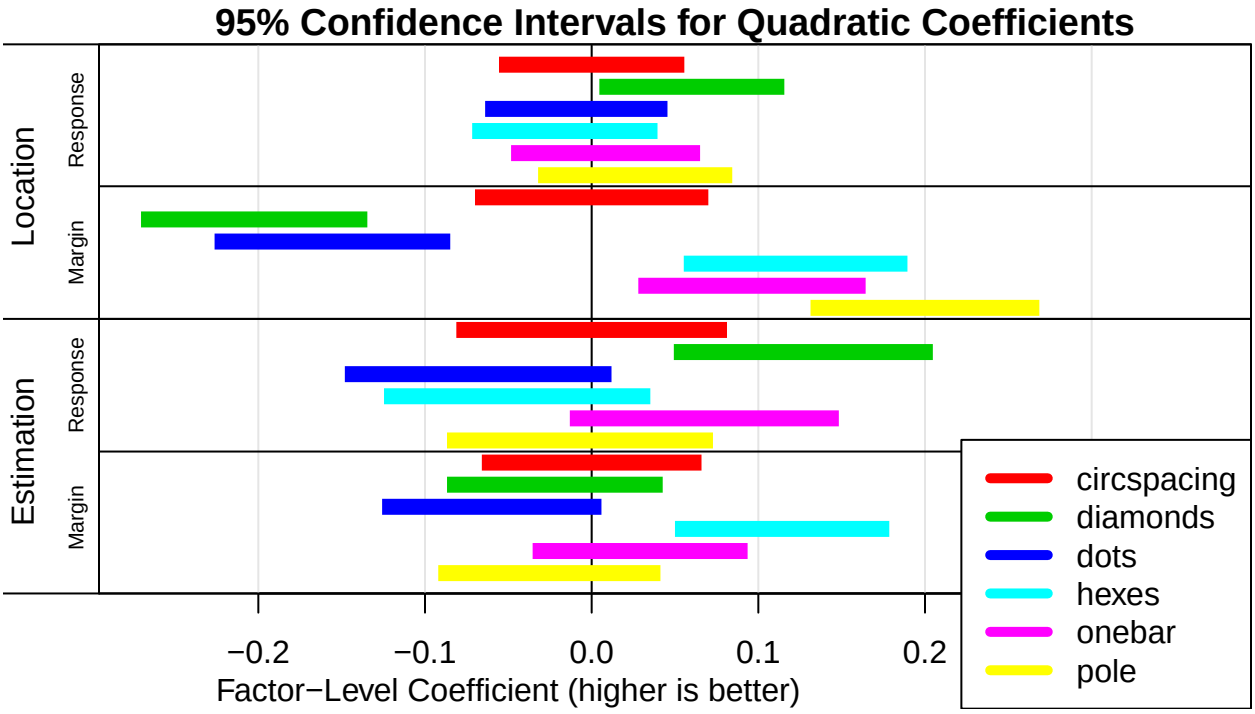


Figure 4.8: Confidence intervals for the coefficient estimates in the quadratic-based models.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
Conclusion

Tables 4.2 and 4.3 show there are interesting patterns in the relationships between `seq` and the response variables. An experiment using human participants is always prone to subjectivity and a relatively high amount of background variability, and this is certainly no exception. However, patterns can be detected and inferences made.

5.1 Interpretation of Significance

The linear regression models suggest that the time to locate or estimate values of desired countries is highly subject to a learning curve, whereas the the error for that location or estimation does not. The participants were essentially given as much time as they needed, so if they needed to take 60 seconds to satisfy their own level of "enough understanding," then the learning curve would be reflected in the `time` variable and not in the `error`. Said another way, they knowingly took more time in order to achieve a comfortable personal confidence level for their selection. Had the experiment instead forced them to respond within 12 seconds, for example, the data might show more significance in the analysis of `error`.

Intuitively, the reader might presume that there would be a different learning curve between questions that ask for the **response**—which utilizes a well-known technique—and questions that ask for the **margin of error**—which uses the newly-introduced techniques. The learning curve for looking at the **response** is possibly related to having to mentally look "through" the secondary technique in order to interpret the color underneath, essentially trying to disregard the distraction or obscuration of the **margin of error**. Table 4.3 suggests that the LOCATION tests show a linear vice quadratic relationship, supporting this intuition, whereas ESTIMATION tests require more of a non-trivial learning process to interpret. In the former, the participant merely has to learn how to focus on the color and not learning about the color itself. In the latter, he or she must not only try to disregard the (possibly distracting) color of the **response**, but might also have to concentrate on interpreting the new shapes.

As stated in Section 4.1, we believe that we do not have independence between a learning curve in the LOCATION test and in the ESTIMATION test: put simply, the participant has spent the last ten minutes already becoming acquainted with the new techniques. If the second half of the test were cognitively as simple as the first half where all he or she had to do was locate a country, then

the learning curve would almost certainly have been shallower and perhaps more linear. However, we compound the dependence between the two by introducing a new task: estimating the value of the shapes. This takes considerably more understanding of the relationship between shape size, spacing, or angle, in order to adequately understand what the plot is telling you. (We might compare this additional mental load to any new task: seeing and recognizing it are one thing, whereas actually doing it or teaching it requires a much deeper understanding of the material.)

Our assertion that a quadratic transformation on `seq` is supported, as shown in Table 4.6 and Figures 4.7 and 4.8. That is, the differences between the two transformations are minor, evidencing more in small shifts in location than with any significant change in standard errors of the coefficients.

5.2 Subjective Participant Responses

After each participant finished the test, they often provided their opinions on individual techniques or on the interface as a whole. Several people suggested, for instance, that we provide to them more information about the pole technique, so that they would know what the black ticks indicated numerically. This was certainly good feedback, but what is interesting is that the information was provided from the beginning, telling the participant that each tick represented 5% increase in the **margin of error**. Each time each technique was used, the left third of the computer screen included text describing that individual technique. The relevance of this is that most people did not read the instructions, they just jumped into the test and performed relatively well. That may speak to the relative intuitiveness of the interface and possibly the techniques.

5.3 Technique Roll-up

The original question, somewhat briefly, is: "Which technique is best?" The answer is: "It depends," mostly on the goal of the plot. Figure 5.1 shows relative performance in four different scenarios, where the closer a dot is to the lower-left corner of the axes, the better overall the technique performs.

If the presenter needs the observers to focus on the performance of the primary variable and tangentially recognize extreme values of the secondary variable, then "Recognize Primary Variable" (lower right) shows that diamonds is the best performing. In this case, error in estimation is not a problem, and aesthetically it could be argued that the colors do not interfere too much with the plot, even though there was evidence to show that it caused slightly more inaccuracy in interpreting the primary variable.

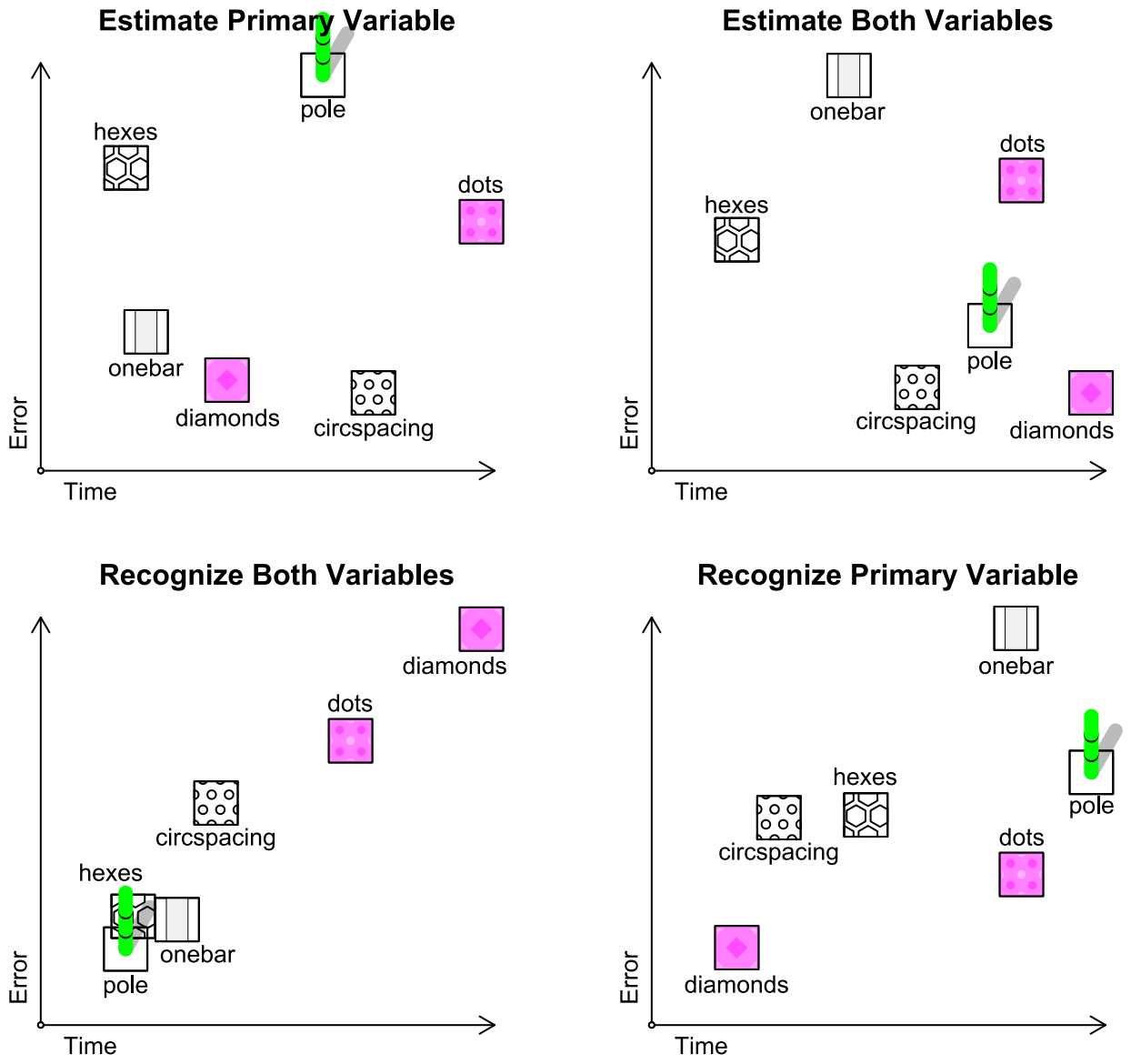


Figure 5.1: Relative performance of all of the techniques in various situations.

"Estimate Both Variables" (upper right) is more appropriate when the observers need to try to quantify the values in various regions in the plot. This is the hardest category in which to rank the techniques, as you can get either "easy" (less time) or "effective" (less inaccurate), but not both.

The other two corners provide slightly different insight into the performance of the techniques. In many cases, a technique is perfectly suited for one situation (e.g., diamonds in the lower right) and the worst performer in others (e.g., diamonds in the lower left).

5.4 Future Work

The design of experiments for varying and controlling factors in the experiment presented numerous options and insight but also considerable problems with a relatively smaller sample-size. For instance, the use of the interaction `var` $\times$ `minmax` might be useful to determine optical illusions where countries that are highly-obscured with the secondary variable are more or less likely to be connected with an extreme in the **response**. Even simpler, `minmax` did not weigh in as a strongly-significant factor in most of the test, but contributed variability nonetheless.

Similarly, requiring the observer to alternate randomly between looking at the **response** and **margin of error**—as well as alternating randomly between other two-level factors—created a frequent and easy opportunity for confusion, evidenced by our assertions of swapped variables in Section 4.1.3, Data Cleaning. Limiting the exposure of the participant to just one variable, or perhaps one variable for long stretches, would minimize the potential for confusion based solely on this factor, an admittedly unimportant effect in the experiment.

We focused heavily in this thesis on the effects of a possible learning curve, and how we could control for its variability in the data. Another piece of the sequencing that we did not analyze in this paper is the sequence within-technique instead of sequence within the test as a whole. For instance, if on a participant's fifth question we present the hexes technique for the first time, then that observation's `seq` value is 5 but its `seq.tech` value is 1. This would provide insight into different learning curves per technique (either intercept-only or intercept and slope). Unfortunately, the *n* of this experiment did not provide enough power or degrees of freedom to subset the data this much and still derive useful statistics.

Another way that the learning curve can be controlled, measured, or even isolated (see Section 4.2 for three suggested types of learning curves) would be to provide a pre-training period. Though the techniques to be tested should not be used in this training, the mechanics of the test and the interface could be countered, as well as having to "look through" one variable to see the other.

After participants finished the experiment, they would typically provide feedback for the techniques or the interface itself. Because we did not track personally-identifiable information in the experiment, it is impossible to see if what they observed to be difficult actually performed differ-

ently; this relationship between perception and reality can be measured with, for example, a survey, where participants can rate each technique and how they feel they did with each technique.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A:
Univariate Plots

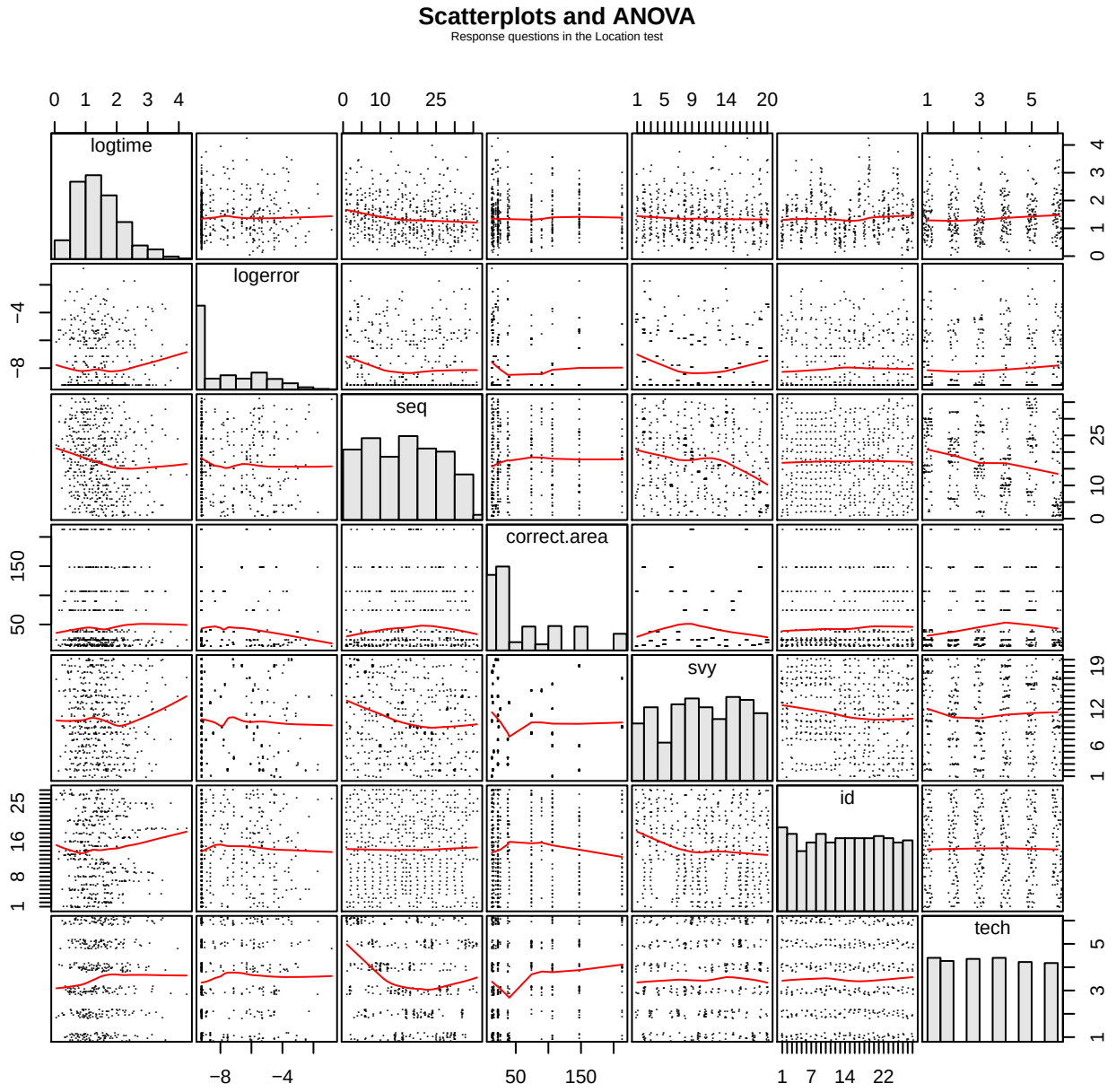


Figure A.1: Pairs plot for `response` questions in the LOCATION test. The `time` and `error` variables have been transformed with a logarithm to help in visualization. The pairs plot for the `margin` questions can be found in Figure 4.4.

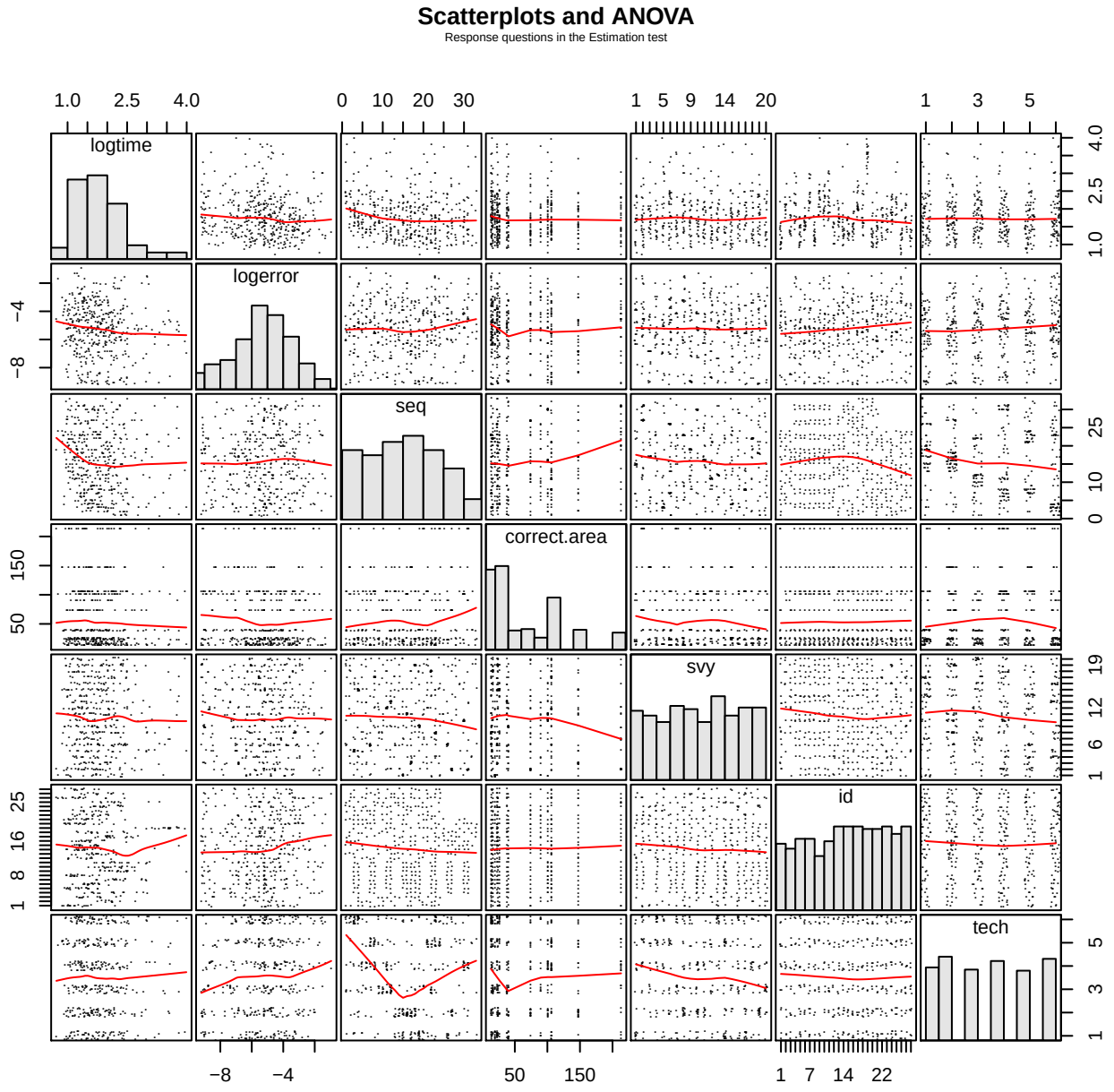


Figure A.2: Pairs plot for `response` questions in the ESTIMATION test. The `time` and `error` variables have been transformed with a logarithm to help in visualization.

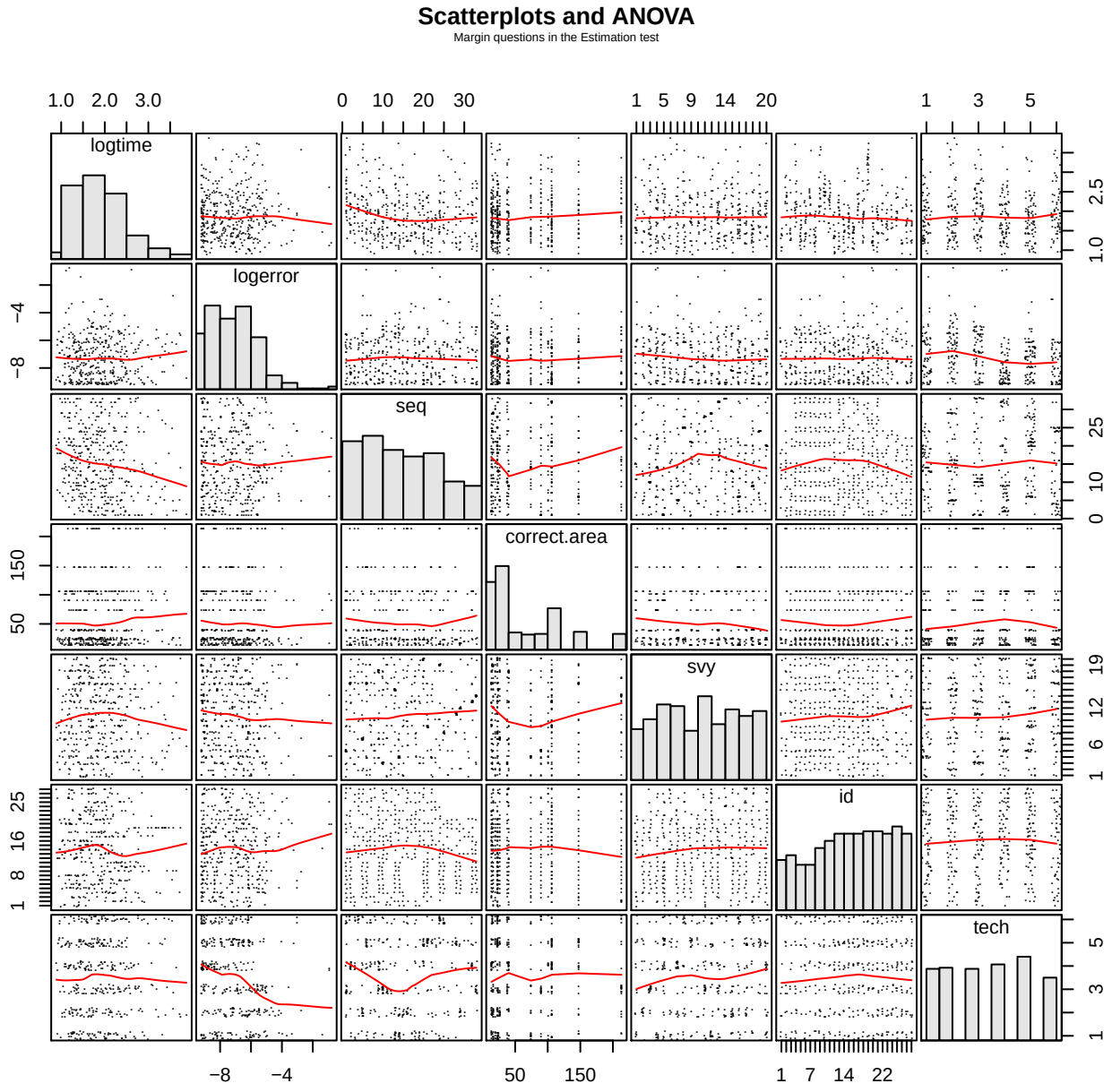


Figure A.3: Pairs plot for `margin` questions in the ESTIMATION test. The `time` and `error` variables have been transformed with a logarithm to help in visualization.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B:
Regression Model Output

Listing B.1: Linear regression of `time` in the LOCATION test, `margin` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity; the range of the removed p-values is listed on the replaced row.

```
Call:
lm(formula = I(log(time)) ~ . - error - RV2 - seq.tech - correct.RESPONSE -
    click.RESPONSE - correct.area, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23785 -0.26516  0.01044  0.25276  1.22857

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.7946908  0.7306043   1.088 0.277334
svy2            0.1025250  0.1305236   0.785 0.432602
...svy...
svy20           0.0838265  0.1507742   0.556 0.578520
id2             0.1165804  0.1549244   0.752 0.452166
...id...
id28            0.3957488  0.1588036   2.492 0.013079 *
seq            -0.0074694  0.0022009  -3.394 0.000754 ***
techdiamonds    0.1833739  0.0754401   2.431 0.015480 *
...tech...
techpole       -0.0340301  0.0743998  -0.457 0.647620
minmaxSMALLEST  0.7615218  0.6263296   1.216 0.224714
click.MARGIN   -1.5906871  0.9581762  -1.660 0.097625 .
click.area     -0.0020397  0.0003561  -5.727 1.93e-08 ***
correct.MARGIN  4.9158927  3.0113649   1.632 0.103323
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.429 on 427 degrees of freedom
Multiple R-squared:  0.6714,    Adjusted R-squared:  0.6284
F-statistic: 15.58 on 56 and 427 DF,  p-value: < 2.2e-16
```

Listing B.2: Linear regression of `error` in the LOCATION test, `margin` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity; the range of the removed p-values is listed on the replaced row.

```
Call:
lm(formula = I(log(1e-04 + error)) ~ . - time - RV2 - seq.tech -
```

```
    correct.area - click.area, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4177 -1.5531 -0.3616  1.3948  6.6647

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -17.03520    4.04900  -4.207 3.15e-05 ***
svy2               1.24394    0.66519   1.870 0.062163 .
...svy...               NA         NA   0.000 0.607305
svy20              2.16148    0.88738   2.436 0.015268 *
id2               -0.45791    0.76808  -0.596 0.551378
...id...                NA         NA   0.031 0.964443
id28              -1.76882    0.78569  -2.251 0.024876 *
seq                0.01941    0.01090   1.781 0.075604 .
techdiamonds       1.71492    0.37438   4.581 6.09e-06 ***
...tech...              NA         NA   0.610 0.773928
techpole          -0.56513    0.37049  -1.525 0.127918
minmaxSMALLEST     6.16861    3.36076   1.835 0.067131 .
click.RESPONSE     0.92323    0.42261   2.185 0.029462 *
click.MARGIN     -16.09617    4.77965  -3.368 0.000827 ***
correct.MARGIN    44.13617   15.87819   2.780 0.005682 **
correct.RESPONSE   1.39252    0.73811   1.887 0.059892 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.121 on 426 degrees of freedom
Multiple R-squared:  0.3125,    Adjusted R-squared:  0.2205
F-statistic: 3.397 on 57 and 426 DF,  p-value: 4.22e-13
```

Listing B.3: Linear regression of `time` in the ESTIMATION test, `response` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity; the range of the removed p-values is listed on the replaced row.

```
Call:
lm(formula = I(log(time)) ~ . - error - RV2 - seq.tech - correct.area -
    correct.RESPONSE - estimate, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67167 -0.21896 -0.02594  0.17462  1.83656

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.338464   0.132681  10.088  < 2e-16 ***
svy2            0.079382   0.113699   0.698   0.4855
...svy...
svy20           0.039614   0.119970   0.330   0.7414
id2             0.191769   0.144654   1.326   0.1857
```

```
...id...
id28             0.213588   0.127902   1.670   0.0958 .
seq             -0.011667   0.002339  -4.988 9.39e-07 ***
techdiamonds    -0.119755   0.065690  -1.823   0.0691 .
...tech...
techpole        -0.028259   0.066020  -0.428   0.6689
correct.MARGIN   0.469014   0.297648   1.576   0.1159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3562 on 373 degrees of freedom
Multiple R-squared:  0.7007,    Adjusted R-squared:  0.6582
F-statistic: 16.48 on 53 and 373 DF,  p-value: < 2.2e-16
```

Listing B.4: Linear regression of `time` in the ESTIMATION test, `margin` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity; the range of the removed p-values is listed on the replaced row.

```
Call:
lm(formula = I(log(time)) ~ . - error - RV2 - seq.tech - estimate -
    correct.RESPONSE, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96595 -0.25818 -0.02995  0.21906  1.62991

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.3351596  0.1886553   7.077 8.34e-12 ***
svy2            0.0670414  0.1543894   0.434  0.66439
...svy...
svy20           0.0093744  0.1354032   0.069  0.94484
id2             0.0349035  0.1872979   0.186  0.85228
...id...
id28            0.5028317  0.1782509   2.821  0.00507 **
seq            -0.0137489  0.0025581  -5.375 1.42e-07 ***
techdiamonds    0.0615356  0.0779618   0.789  0.43048
...tech...
techpole        0.1024772  0.0799172   1.282  0.20061
correct.MARGIN  0.6313386  0.3589594   1.759  0.07950 .
correct.area    0.0005766  0.0004183   1.378  0.16896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4059 on 343 degrees of freedom
Multiple R-squared:  0.6106,    Adjusted R-squared:  0.5493
F-statistic:  9.96 on 54 and 343 DF,  p-value: < 2.2e-16
```

Listing B.5: Linear regression of `error` in the ESTIMATION test, `response` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity; the range of the removed p-values is listed on the replaced row.

```
Call:
lm(formula = I(log(1e-04 + error)) ~ . - time - RV2 - seq.tech -
    seq - correct.MARGIN - correct.area, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3352 -0.9103  0.1191  1.0239  5.0067

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.1146     0.6492  -9.419  < 2e-16 ***
svy2              -0.9253     0.5272  -1.755  0.08005 .
...svy...              NA         NA   0.037  0.96616
svy20             -0.2013     0.5694  -0.353  0.72393
id2                0.5358     0.6722   0.797  0.42588
...id...               NA         NA   0.026  0.99474
id28              -0.1117     0.5956  -0.187  0.85138
techdiamonds       0.1528     0.3066   0.498  0.61849
...tech...             NA         NA   0.147  0.19236
techpole           0.8692     0.3002   2.895  0.00402 **
estimate          -3.8901     0.6680  -5.823 1.25e-08 ***
correct.RESPONSE   5.2297     0.6783   7.710 1.15e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.656 on 373 degrees of freedom
Multiple R-squared:  0.3173,    Adjusted R-squared:  0.2203
F-statistic: 3.271 on 53 and 373 DF,  p-value: 1.904e-11
```

Listing B.6: Linear regression of `error` in the ESTIMATION test, `margin` questions. Display of the `id`, `svy`, and `tech` indicator variables are abbreviated for brevity; the range of the removed p-values is listed on the replaced row.

```
Call:
lm(formula = I(log(1e-04 + error)) ~ . - time - RV2 - seq.tech -
    correct.area - seq - correct.RESPONSE, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4863 -0.8777 -0.0607  0.8393  5.5851

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.659554   0.599861 -11.102  < 2e-16 ***
```

```
svy2            -0.633453   0.494024  -1.282    0.2006
...svy...              NA         NA   0.084    0.9645
svy20           -0.641302   0.435344  -1.473    0.1416
id2             -0.226443   0.606163  -0.374    0.7090
...id...               NA         NA   0.117    0.9704
id28             0.001756   0.574113   0.003    0.9976
techdiamonds     0.585031   0.252368   2.318    0.0210 *
...tech...             NA         NA   0.000    0.6780
techpole        -0.507206   0.256990  -1.974    0.0492 *
estimate         8.794849   0.994351   8.845   < 2e-16 ***
correct.MARGIN  -9.228746   1.464384  -6.302 8.99e-10 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.313 on 344 degrees of freedom
Multiple R-squared:  0.4065,    Adjusted R-squared:  0.3151
F-statistic: 4.445 on 53 and 344 DF,  p-value: < 2.2e-16
```

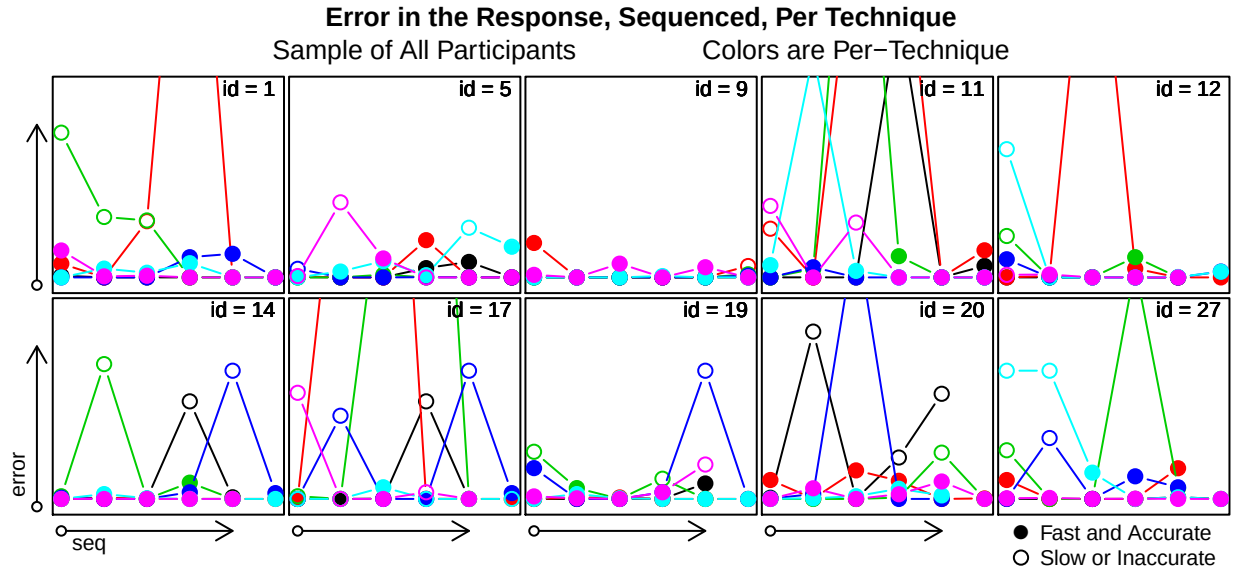


Figure B.1: Error per sequence, colored by technique and paneled by participant. The important point to take from these plots is the difference in variability in `error` between participants.

THIS PAGE INTENTIONALLY LEFT BLANK

References

Bates, Douglas, Martin Maechler, Ben Bolker. 2013. *lme4: Linear mixed-effects models using S4 classes*. URL `http://CRAN.R-project.org/package=lme4`. R package version 0.999999-2.

Cleveland, William S. 1993. *Visualizing Data*. Hobart Press, Summit, NJ.

Cleveland, William S. 1994. *The Elements of Graphing Data*. Hobart Press, Summit, NJ.

Cleveland, William S., Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical models. *Journal of the American Statistical Association* **79**(387) 531–554.

Cognitive. 2013. In *Merriam-Webster*.com. Retrieved Aug 14, 2013, from `http://www.merriam-webster.com/dictionary/cognitive`.

Faraway, Julian J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Taylor and Francis Group, Boca Raton, FL.

Fricker, Ronald, Samuel Buttrey, William Evans. In press. Visualization for sociocultural signature detection. Jill Egeth, Gary Klein, Dylan Schmorrow, eds., *Sociocultural Behavior Sensemaking: State of the Art in Understanding the Operational Environment*. The MITRE Corporation, unknown pages.

Hsu, Fred. 2005. Stereogram Tut Shark Depthmap.png. Retrieved on Aug 25, 2013, from `http://en.wikipedia.org/wiki/File:Stereogram_Tut_Shark_Depthmap.png`.

Hyman, Ray. 1953. Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology* **45** 188–196.

Intuition. 2013. In *Merriam-Webster*.com. Retrieved Aug 14, 2013, from `http://www.merriam-webster.com/dictionary/intuition`.

Livingston, Mark A., Jonathan Decker, Zhuming Ai. 2011. An evaluation of methods for encoding multiple, 2d spatial data. *SPIE Visualization and Data Analysis* Burlingame, CA.

Livingston, Mark A., Jonathan W. Decker. 2011. Evaluation of trend localization with multi-variate visualization. *IEEE Transactions on Visualization and Computer Graphics* **17**(12) 2053–2062.

Livingston, Mark A., Jonathan W. Decker. 2012. Evaluation of multi-variate visualizations: A case study of refinements and user experience. *SPIE Visualization and Data Analysis* Burlingame, CA.

Livingston, Mark A., Jonathan W. Decker, Zhuming Ai. 2012. Evaluation of multivariate visualization on a multivariate task. *IEEE Transactions on Visualization and Computer Graphics* **18**(12) 2114–2121.

Livingston, Mark A., Jonathan W. Decker, Zhuming Ai. 2013. Evaluating multivariate visualizations on time-varying data. *Proceedings of SPIE Visualization and Data Analysis* **8654**. Burlingame, CA.

National Institute of Standards and Technology. 2012. What is EDA? Retrieved Aug 14, 2013, from `http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm`.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Rosling, H. 2013. Gapminder world. Retrieved Apr 20, 2013, from `http://www.gapminder.org`.

StackOverflow user JT85. 2013. Plots on a map using ggplot2. Retrieved Aug 14, 2013, from `http://stackoverflow.com/questions/16028659/plots-on-a-map-using-ggplot2`.

Sviokla, John. 2009. Swimming in data? Three benefits of visualization. Retrieved Apr 16, 2013, from `http://blogs.hbr.org/sviokla/2009/12/swimming_in_data_three_benefit.html`.

Tufte, Edward Rolf. 1990. *Envisioning Information*. Graphics Press LLC, Cheshire, CT.

Tufte, Edward Rolf. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Press LLC, Cheshire, CT.

Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Wickham, Hadley. 2009. *ggplot2, Elegant Graphics for Data Analysis*. Springer, New York, NY.

Wickham, Hadley. 2013. Wrap a 1d ribbon of panels into 2d. Retrived Aug 14, 2013, from `http://docs.ggplot2.org/current/facet_wrap.html`.

Yau, Nathan. 2011. *Visualize This, The Flowingdata Guide to Design, Visualization, and Statistics*. Wiley Publishing, Inc, Indianapolis, IN.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia
2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California